High Dimensional Statistical Estimation under Uniformly Dithered One-bit Quantization*

Junren Chen[1,*], Cheng-Long Wang[2,†], Michael K. Ng[1,‡], Di Wang[2,3,‡]

August 19, 2022

Abstract

In this paper, we propose a uniformly dithered one-bit quantization scheme for high-dimensional statistical estimation. The scheme contains truncation, dithering, and quantization as typical steps. As canonical examples, the quantization scheme is applied to three estimation problems: sparse covariance matrix estimation, sparse linear regression, and matrix completion. We study both sub-Gaussian and heavy-tailed regimes, with the underlying distribution of heavy-tailed data assumed to possess bounded second or fourth moment. For each model we propose new estimators based on one-bit quantized data. In sub-Gaussian regime, our estimators achieve optimal minimax rates up to logarithmic factors, which indicates that our quantization scheme nearly introduces no additional cost. In heavy-tailed regime, while the rates of our estimators become essentially slower, these results are either the first ones in such one-bit quantized and heavy-tailed setting, or exhibit significant improvements over existing comparable results. Moreover, we contribute considerably to the problems of one-bit compressed sensing and one-bit matrix completion. Specifically, we extend one-bit compressed sensing to sub-Gaussian or even heavy-tailed sensing vectors via convex programming. For one-bit matrix completion, our method is essentially different from the standard likelihood approach and can handle pre-quantization random noise with unknown distribution. Experimental results on synthetic data are presented to support our theoretical analysis.

*Junren Chen and Michael K. Ng were supported in part by Hong Kong Research Grant Council GRF 12300218, 12300519, 17201020, 17300021, C1013-21GF, C7004-21GF and Joint NSFC-RGC N-HKU76921. Di Wang and Cheng-Long Wang were supported in part by the baseline funding BAS/1/1689-01-01, funding from the CRG grand URF/1/4663-01-01, FCC/1/1976-49-01 from CBRC and funding from the AI Initiative REI/1/4811-10-01 of King Abdullah University of Science and Technology (KAUST).

[1]Department of Mathematics, The University of Hong Kong, Pokfulam, Hong Kong.

[2]Division of CEMSE, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia.

[3]Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia.

[*]Email: chenjr58@connect.hku.hk

[†]Email: chenglong.wang@kaust.edu.sa

[‡]Corresponding author. Email: mng@maths.hku.hk, di.wang@kaust.edu.sa

1 Introduction

One-bit quantization of signals or data recently has received much attention in both signal processing and machine learning communities. In some signal processing problems, power consumption, manufacturing cost and chip area of analog-to-digital devices grow exponentially with their resolution [54], thus, it is impractical and infeasible to use high-precision data or signals. Alternatively, it was proposed to use low-resolution, specifically one-bit quantization, see for instance [4, 31, 34, 53, 64, 77]. In a more general sense, the quantization process that maps an analog signal into digital representation of a finite dictionary is a critical process in signal processing. Besides, in many distributed machine learning or federated learning scenarios, multiple parties transmit information among themselves. The communication cost can be prohibitive for distributed algorithms where each party only has a low-power and low-bandwidth device such as a mobile device [60]. To address the bottleneck of communication cost, recent works have studied how to send a small number or even one bit per entry for such distributed machine learning applications [3, 8, 79, 86]. However, theoretical results on statistical estimation under data quantization are extremely rare in literature, hence the behaviour of a learning process based on quantized data remains unclear. Recently, Dirksen et al. [36] proposed a one-bit estimator for an unstructured covariance matrix. Although their estimator can achieve a near minimax rate, their method is only for low dimensional case and cannot be extended to the high dimensional (approximately) sparse or low rank setting, which is commonly encountered in statistics and signal processing.

We further fill the severe theoretical vacancy in this paper. More specifically, we study three fundamental high-dimensional statistical estimation problems based on data that are quantized to one bit. The quantization scheme include the typical steps of truncation, dithering, and quantization (note that truncation is for heavy-tailed data only, which shrinks outliers and enhances robustness), see Section 1.2 for detailed discussions. We present extensive theoretical results on sparse covariance matrix estimation, sparse linear regression, and low-rank matrix completion, under both sub-Gaussian data and heavy-tailed data. Here, the underlying distribution of heavy-tailed data is assumed to have bounded second or fourth moment. Our estimators in sub-Gaussian regime have remarkable statistical properties, i.e., they could achieve near minimax rates (up to some logarithmic factors). In the heavy-tailed regime, our estimators can still deliver a faithful estimation under a high-dimensional scaling, while due to a bias-and-variance trade-off, the error rates are essentially slower than the minimax ones in the classical settings. However, to our best knowledge, these are the first high-dimensional statistical results under such two-fold predicament, i.e., heavy-tailed distribution that breaks the robustness, and one-bit quantization that loses data information. Here we summarize our key results and contributions as follows (For simplicity we only consider parameters n, d, s (or r), q and omit the others).

- In Section 2, for some zero-mean d-dimensional random vector X, we study the problem of estimating its covariance matrix $\mathbf{\Sigma}^* = \mathbb{E}\big(XX^T\big) = [\sigma^*_{ij}]$, where $\mathbf{\Sigma}^*$ has the approximate column-wise sparsity structure, i.e., $\sup_{j\in[d]}\sum_{i=1}^d |\sigma^*_{ij}|^q \leq s$ for some $0 \leq q < 1$ and $s > 0$. Denote the full data that are i.i.d. copies of X by $X_1, ..., X_n$. For sub-Gaussian X, we i.i.d. sample the dithering noise vector $\{\Gamma_{k1}, \Gamma_{k2} : k \in [n]\}$ that are uniformly distributed on $[-\gamma, \gamma]^d$, and then dither and quantize each X_k to binary data $\text{sign}(X_k + \Gamma_{k1}),\ \text{sign}(X_k + \Gamma_{k2})$. Based on these binary data, we propose a thresholding estimator $\widehat{\mathbf{\Sigma}}$, see (2.4) and (2.18). Although only two bits are collected per entry, we show a near optimal minimax rate

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\mathrm{op}} \lesssim s \log n \Big(\frac{\log d}{n}\Big)^{(1-q)/2}.$$

For heavy-tailed X assumed to have bounded fourth moment, we first element-wisely truncate the full sample X_k to be $\widetilde{X}_k := \mathrm{sign}(X_k)\min\{|X_k|, \eta\}$ (element-wise operation). Then similar to sub-Gaussian data, we deal with $\widetilde{X}_k$ by dithering and quantization. Our estimator possesses an estimation error bound for operator norm error

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\mathrm{op}} \lesssim s \Big(\frac{\log d}{n}\Big)^{(1-q)/4}.$$

- In Section 3, we study sparse linear regression $Y_k = X_k^T\Theta^* + \epsilon_k,\ k \in [n]$ where the desired signal $\Theta^* = [\theta_i^*] \in \mathbb{R}^d$ satisfies $\sum_{i=1}^d |\theta_i^*|^q \leq s$ for some $0 \leq q < 1$ and $s > 0$, the covariate X_k and the additive noise ϵ_k can be either sub-Gaussian or heavy-tailed. Given the full data $\{(X_k, Y_k) : k \in [n]\}$, we first study a novel setting where both X_k and Y_k are quantized to binary data. The covariate X_k is quantized by exactly the same method as Section 2. For sub-Gaussian X_k and ϵ_k, the response Y_k is quantized to be $\mathrm{sign}(Y_k + \Lambda_k)$ with Λ_k uniformly distributed on $[-\gamma, \gamma]$. When X_k and ϵ_k are heavy-tailed (with bounded fourth moment), we truncate Y_k to be $\widetilde{Y}_k$ and then similarly apply the dithered quantization to $\widetilde{Y}_k$. The estimation relies on the one-bit sparse covariance matrix estimator $\widehat{\mathbf{\Sigma}}$ developed in Section 2. To deal with the lack of positive semi-definiteness, we assume $\mathbf{\Sigma}_{XX} = \mathbb{E}X_kX_k^T$ has column-wise sparsity, which accommodates the conventional isotropic condition (i.e., $\mathbf{\Sigma}_{XX} = \boldsymbol{I_d}$) used in compressed sensing. We formulate the recovery as a convex programming problem with objective function combining a generalized quadratic loss and an ℓ_1 regularizer, see (3.17). In sub-Gaussian case, we show our estimator $\widehat{\Theta}$ could achieve a near optimal minimax rate of

$$\|\widehat{\Theta} - \Theta^*\|_2 \lesssim \sqrt{s}\Big(\log n\sqrt{\frac{\log d}{n}}\Big)^{1-q/2}.$$

In heavy-tailed case, our estimator possesses the error rate

$$\|\widehat{\Theta} - \Theta^*\|_2 \lesssim \sqrt{s}\Big(\frac{\log d}{n}\Big)^{(1-q/2)/4}.$$

Besides the first results for this new setting, we also revisit the canonical one-bit compressed sensing problem where we quantize Y_k in a same manner but have full knowledge of X_k. We estimate Θ^* via analogous convex programming problems, see (3.24) and (3.28). In sub-Gaussian regime, our estimator achieves a near optimal minimax rate

$$\|\widehat{\Theta} - \Theta^*\|_2 \lesssim \sqrt{s}\Big(\sqrt{\frac{\log d \log n}{n}}\Big)^{1-q/2}.$$

In heavy-tailed regime, our estimator can still handle the high-dimensional sparse setting which satisfies

$$\|\widehat{\Theta} - \Theta^*\|_2 \lesssim s^{\frac{2}{3}}\Big(\frac{\log d}{n}\Big)^{(1-q/2)/3}.$$

As it turns out, these two results exhibit significant improvements upon existing worsk (e.g., more general sensing vector, recovery from convex programming, or faster error rate), see a detailed comparison in Appendix D.

- In Section 4, we study the problem of low-rank matrix completion $Y_k = \langle \boldsymbol{X_k}, \boldsymbol{\Theta}^* \rangle + \epsilon_k$, where the desired $d \times d$ matrix $\boldsymbol{\Theta}^*$ with singular values $\sigma_1(\boldsymbol{\Theta}^*) \geq ... \geq \sigma_d(\boldsymbol{\Theta}^*)$ is (approximately) low-rank $\sum_{k=1}^d \sigma_k(\boldsymbol{\Theta}^*)^q \leq r$ for some $0 \leq q < 1$ and $r > 0$. The covariate $\boldsymbol{X_k}$ is uniformly distributed on $\{e_i e_j^T : i, j \in [d]\}$ where e_i is the i-th column of the $\boldsymbol{I_d}$, ϵ_k is sub-Gaussian or heavy-tailed noise. Given the full data $\{(\boldsymbol{X_k}, Y_k)\}$, we quantize Y_k to one bit by the same process as one-bit compressed sensing in Section 3. Our estimator $\widehat{\boldsymbol{\Theta}}$ is given by minimizing an objective functions constituted of a generalized quadratic loss and a nuclear norm penalty, see (4.6). If ϵ_k is sub-Gaussian, we show that $\widehat{\boldsymbol{\Theta}}$ achieves a near optimal minimax rate

$$\frac{\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2}{d^2} \lesssim rd^{-q}\Big(\log n \frac{d \log d}{n}\Big)^{1-q/2},$$

  If ϵ_k is heavy-tailed with bounded second moment, we show the recovery guarantee

$$\frac{\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2}{d^2} \lesssim rd^{-q}\Big(\frac{d \log d}{n}\Big)^{1/2-q/4}.$$

  We emphasize that our approach is totally different from the only existing method for one-bit matrix completion, i.e., based on maximizing a likelihood function. Our essential advantage is that our method can handle unknown pre-quantization random noise that can even be heavy-tailed. See more discussions in Appendix D.

The rest of the paper is given as follows. In Section 5 we provide an overview of the proofs and the main techniques; In Section 6, we present experimental results to corroborate our theories; We finally give some concluding remarks in Section 7. The complete proofs, the comparisons with existing results and details of numerical experiments are deferred to the Appendix.

1.1 Notations and Preliminaries

As general principles, lowercase letters (e.g., s, r) represent scalars, capital letters (e.g., X, Y) represent vectors, and capital bold letters (e.g., $\boldsymbol{X}, \boldsymbol{\Theta}$) represent matrices. Some exceptions are that we use capital letter Y, Y_k to denote the responses, Λ, Λ_k to denote the dithering noise for Y, Y_k, and $X_{k,i}$ for the i-th entry of X_k. Notations marked by $*$ denote the desired underlying signals, e.g., $\boldsymbol{\Sigma}^*, \Theta^*, \boldsymbol{\Theta}^*$, while those with a hat denote our estimators, e.g., $\widehat{\boldsymbol{\Sigma}}, \widehat{\Theta}, \widehat{\boldsymbol{\Theta}}$.

We first introduce different vector or matrix norms. Let $[N] = \{1, 2, ..., N\}$. For a vector $X = [x_i] \in \mathbb{R}^d$, the ℓ_1 norm, ℓ_2 norm and max norm are given by $\|X\|_1 = \sum_{i=1}^d |x_i|$, $\|X\|_2 = (\sum_{i=1}^d |x_i|^2)^{1/2}$, $\|X\|_{\max} = \max_{i \in [d]} |x_i|$, respectively. Note that we also use $\|X\|_0$ to denote the number of non-zero entries in X. For a matrix $\boldsymbol{X} = [x_{ij}] \in \mathbb{R}^{d \times d}$, the operator norm, Frobenius norm and max norm are defined as $\|\boldsymbol{X}\|_{\mathrm{op}} = \sup_{\|V\|_2 = 1} \|\boldsymbol{X}V\|_2$, $\|\boldsymbol{X}\|_{\mathrm{F}} = (\sum_{i=1}^d \sum_{j=1}^d x_{ij}^2)^{1/2}$, $\|\boldsymbol{X}\|_{\max} = \max_{1 \leq i,j \leq d} |x_{ij}|$. Assume the singular values are $\sigma_1(\boldsymbol{X}) \geq \sigma_2(\boldsymbol{X}) \geq ... \geq \sigma_d(\boldsymbol{X})$, then the nuclear norm $\|\boldsymbol{X}\|_{\mathrm{nu}} = \sum_{i=1}^d \sigma_i(\boldsymbol{X})$ serves as the counterpart of the ℓ_1 norm of vectors. Given $\boldsymbol{A} = [\alpha_1, ..., \alpha_d] \in \mathbb{R}^{d \times d}$, we use $\mathrm{vec}(\cdot)$ to vectorize $\boldsymbol{A}$ in a column-wise manner, i.e., $\mathrm{vec}(A) = [\alpha_1^T, \alpha_2^T, ..., \alpha_d^T]^T$, while the inverse of $\mathrm{vec}(\cdot)$ is denoted by $\mathrm{mat}(\cdot)$. Assume $\boldsymbol{B} \in \mathbb{R}^{d \times d}$, then the inner product in $\mathbb{R}^{d \times d}$ is defined by $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{Tr}(\boldsymbol{A}^T \boldsymbol{B}) = \mathrm{vec}(\boldsymbol{A})^T \mathrm{vec}(\boldsymbol{B})$.

Throughout the paper, we use n to denote the number of samples in data, d to denote the dimension of the feature vector of each sample. Expectation and probability are denoted by $\mathbb{E}(\cdot)$, $\mathbb{P}(\cdot)$ respectively. For a specific event E, $\mathbb{1}(E)$ stands for the corresponding indicator function, i.e., $\mathbb{1}(E) = 1$ if E happens, $\mathbb{1}(E) = 0$ otherwise. We work with quite a lot of parameters arising in several signal processing steps. To avoid confusion of constants, we use $\{D_1, D_2, D_3, ...\}$ to denote constants whose values may vary from line to line, while $\{C_1, C_2, C_3, ...\}$ would only be used once to set a specific parameter, see (2.6), (2.9) for example.

We adopt standard asymptotic notations that omits absolute constants. Specifically, we use $B_1 \lesssim B_2$ or $B_1 = O(B_2)$ to abbreviate the fact that $B_1 \leq CB_2$ for some absolute constant C. Similarly, we write $B_1 \gtrsim B_2$ or alternatively $B_1 = \Omega(B_2)$ if $B_1 \geq CB_2$ for some $C > 0$. If both $B_1 = O(B_2)$ and $B_1 = \Omega(B_2)$ hold, i.e., B_1 equals B_2 up to constants, we write $B_1 \asymp B_2$.

The function $\text{sign}(\cdot)$ extracts the sign of a real number x, i.e., $\text{sign}(x) = 1$ if $x \geq 0$, $\text{sign}(x) = -1$ if $x < 0$. Hard thresholding operator with threshold ζ is defined by $\mathcal{T}_\zeta(x) = x\mathbb{1}(|x| \geq \zeta)$. Both $\text{sign}(\cdot)$ and $\mathcal{T}_\zeta(\cdot)$ operate on vectors or matrices element-wisely.

To broaden the range of our readers, we give some preliminaries on sub-Gaussian random variable or concentration inequality as follows.

Definition 1. *Given a real random variable $X \in \mathbb{R}$, its sub-Gaussian norm $\|X\|_{\psi_2}$, sub-exponential norm $\|X\|_{\psi_1}$ are defined as*

$$\|X\|_{\psi_2} = \inf\Big\{t > 0 : \mathbb{E}\exp\Big(\frac{X^2}{t^2}\Big) \leq 2\Big\} \quad , \quad \|X\|_{\psi_1} = \inf\Big\{t > 0 : \mathbb{E}\exp\Big(\frac{|X|}{t}\Big) \leq 2\Big\}. \tag{1.1}$$

X is said to be sub-Gaussian if $\|X\|_{\psi_2} \leq \infty$.

Definition 2. *Given a real random vector $X \in \mathbb{R}^d$, the sub-Gaussian norm is defined to be $\|X\|_{\psi_2} = \sup_{\|V\|_2=1} \|V^T X\|_{\psi_2}$. X is said to be sub-Gaussian if $\|X\|_{\psi_2} \leq \infty$.*

For $X, Y \in \mathbb{R}$ we note a useful relation (see Lemma 2.7.7 in [87] for instance)

$$\|XY\|_{\psi_2} \leq \|X\|_{\psi_1}\|Y\|_{\psi_1}. \tag{1.2}$$

Sub-Gaussian variable X share similar properties with Gaussian random variable, such as light probability tail and bounded moment constraint. Here we only introduce the properties used in this work, and interesting readers shall refer to [87].

Proposition 1. (Proposition 2.5.2, [87] *Assume random variable X is sub-Gaussian, then for absolute constants D_1, D_2 we have:*

(a) *For any $t > 0$, $\mathbb{P}\big(|X| \geq t\big) \leq 2\exp\big(-\frac{D_1 t^2}{\|X\|_{\psi_2}^2}\big)$.*

(b) *For any $p \geq 1$, $\big(\mathbb{E}|X|^p\big)^{1/p} \leq D_2\|X\|_{\psi_2}\sqrt{p}$.*

Proposition 2. (Proposition 2.6.1, [87]) *Let $X_1, ..., X_N$ be independent, zero-mean, sub-Gaussian random variables, then for some absolute constant D_1 we have $\big\|\sum_{k=1}^N X_k\big\|_{\psi_2}^2 \leq D_1 \sum_{k=1}^N \|X_k\|_{\psi_2}^2$.*

For concentration results, we only introduce Hoeffding's inequality and Bernstein's inequality. Several other concentration inequalities (e.g., Matrix Bernstein's inequality) would be properly referred to the sources when they are invoked in the proof.

Proposition 3. (Hoeffding's inequality, e.g., Theorem 1.9, [76]) *Let $X_1, ..., X_n$ be independent, bounded random variables satisfying $X_i \in [a_i, b_i]$, then for any $t > 0$ it holds that*

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{k=1}^n (X_k - \mathbb{E}X_k)\Big| \geq t\Big) \leq 2\exp\Big(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big). \tag{1.3}$$

Proposition 4. (Bernstein's inequality, e.g., Theorem 2.8.1, [87]) *Let $X_1,...,X_N$ be independent random variables, then for any $t>0$ and for some absolute constant D_1 we have*

$$\mathbb{P}\Big(\big|\sum_{k=1}^{N}(X_k-\mathbb{E}X_k)\big|\geq t\Big)\leq 2\exp\Big(-D_1\min\big\{\frac{t^2}{\sum_{k=1}^N\|X_k\|_{\psi_1}^2},\frac{t}{\max_{k\in[N]}\|X_k\|_{\psi_1}}\big\}\Big). \quad (1.4)$$

Although sub-Gaussian data has exciting statistical properties like similar tail bounds as Gaussian distribution, data in some real problems may have much heavier tail, to name a few, data in economics and finance [49], biomedical data [13,90], noise in signal processing [82,89], and even signal itself [2,61,63]. Therefore, it is necessary to study one-bit statistical estimation in heavy-tailed regime, which is also considered in this work. We use bounded moment of certain order to capture the heavy-tailedness, i.e., $\mathbb{E}|X|^l\leq M$ for some $l\in\mathbb{N}_+$. Note that this is a widely used strategy [42, 43, 48, 52, 81, 89, 92].

1.2 One-bit Quantization Scheme

Truncation, dithering and quantization are three typical signal processing steps in our work. Here we summarize our one-bit quantization scheme briefly:

1. **Truncation.** The truncation step will only be used to heavy-tailed data. Specifically, we first specify a threshold $\eta>0$, then the truncation step shrinks a scalar x to be $\text{sign}(x)\min\{|x|,\eta\}$, and hence x with magnitude smaller than η would remain unchanged in truncation. Vectors are truncated element-wisely, and notations marked by tilde are used exclusively to denote truncated data, for example, $\widetilde{X}_k$ and $\widetilde{Y}_k$.

2. **Dithering.** The dithering step is applied to all the data that we plan to quantize to one bit. For $E\subset\mathbb{R}^m$, we use $X\sim\text{uni}(E)$ to state that X obeys uniform distribution on E. In sub-Gaussian case we dither the covariate X_k and response Y_k by uniformly distributed noise. Specifically, as we need to sample two bits per entry for X_k, we draw $\Gamma_{k1},\Gamma_{k2}\sim\text{uni}([-\gamma,\gamma]^d)$ and dither X_k to be $X_k+\Gamma_{k1}$, $X_k+\Gamma_{k2}$. We only need 1-bit information for each response Y_k, so we sample $\Lambda_k\sim\text{uni}([-\gamma,\gamma])$ and obtain the dithered response $Y_k+\Lambda_k$. In heavy-tailed case X_k and Y_k are substituted by the truncated data $\widetilde{X}_k$ and $\widetilde{Y}_k$.

3. **Quantization.** In quantization step we simply apply $\text{sign}(\cdot)$ to the dithered data, and notations marked by a dot (e.g., $\dot{Y}_k,\dot{X}_{k1},\dot{X}_{k2}$) exclusively represent the one-bit quantized data. More precisely, we have $\dot{Y}_k=\text{sign}(Y_k+\Lambda_k)$, $\dot{X}_{kj}=\text{sign}(X_k+\Gamma_{kj})$, $j=1,2$ for sub-Gaussian X_k, Y_k, and $\dot{Y}_k=\text{sign}(\widetilde{Y}_k+\Lambda_k)$, $\dot{X}_{kj}=\text{sign}(\widetilde{X}_k+\Gamma_{kj})$, $j=1,2$ for heavy-tailed X and Y.

1.3 Intuition and Heuristic

In this subsection we illustrate the main idea of our approaches before proceeding to details. Specifically, we will provide the intuition of our one-bit quantization scheme, and then heuristically analyse a multi-bits matrix completion setting to illustrate the the reason why our estimators could achieve near optimal minimax rates in sub-Gaussian regime. In fact, the idea of the whole paper stems from two simple observations, which are given in the following two lemmas. We mention that Corollary 1 motivates [36] to estimate $\mathbb{E}(XY)$ and hence an unstructured covariance matrix via binary data, while Lemma 1 is its more general form and enlightens the estimators in our work. For instance, while full observations are not available, our loss function in matrix completion is constructed by substituting the full data Y_k in the empirical ℓ_2 loss with the one-bit surrogate $\gamma \cdot \dot{Y}_k$ (see (4.6)). This idea comes from Lemma 1.

Lemma 1. *Let X, Λ be two independent random variables satisfying $|X| \le B$, $\Lambda \sim \text{uni}\big([-\gamma,\gamma]\big)$ where $\gamma \ge B$, then we have $\mathbb{E}\big[\gamma \cdot \text{sign}(X+\Lambda)\big] = \mathbb{E}X$.*

Corollary 1. (Lemma 16, [36]) *Let X, Y be bounded random variables satisfying $|X| \le B$, $|Y| \le B$, Λ_1, Λ_2 are i.i.d. uniformly distributed on $[-\gamma,\gamma]$, $\gamma \ge B$, and Λ_1, Λ_2 are independent of X, Y. Then we have $\mathbb{E}\big[\gamma^2 \cdot \text{sign}(X+\Lambda_1)\cdot\text{sign}(Y+\Lambda_2)\big] = \mathbb{E}XY$.*

Next, by informal arguments, we heuristically compare full-data-based matrix completion and quantized matrix completion where one can sample finitely many bits from each Y_k (namely multi-bits matrix completion). This comparison can provide some insights of why our estimators can achieve a near optimal minimax rate in sub-Gaussian regime.

We consider a full-data sample of size n from matrix completion (4.1) and denote it by

$$\mathcal{D}_{\text{full}} = \Big\{(\boldsymbol{X_1}, Y_1), ..., (\boldsymbol{X_n}, Y_n)\Big\}.$$

Now, for some positive integer $f(n)$ we i.i.d. draw $\{\Lambda_{kj} : j \in [f(n)]\}$ from $\text{uni}\big([-\gamma,\gamma]\big)$, and sample $f(n)$ bits from each Y_k by the proposed dithered quantization, that is, $\{\dot{Y}_{kj} := \text{sign}(Y_k + \Lambda_{kj}) : j \in [f(n)]\}$. This quantization process yields the sample containing $n \cdot f(n)$ binary observations

$$\mathcal{D}_{\text{mult}} = \Big\{(\boldsymbol{X_k}, \dot{Y}_{kj}) : k \in [n],\ j \in [f(n)]\Big\}.$$

Interestingly, from $\mathcal{D}_{\text{mult}}$ one can build a dataset with size n as

$$\mathcal{D}_{\text{appr}} = \left\{(\boldsymbol{X_k}, Y_{k,\text{appr}}) : Y_{k,\text{appr}} = \frac{1}{f(n)} \sum_{j\in[f(n)]} \gamma \cdot \dot{Y}_{kj},\ k \in [n]\right\}.$$

We aim to reveal that the above three samples are comparably informative for the estimation.

For simplicity we assume $\|\epsilon_k\|_{\psi_2} = O(1)$, $\|Y_k\|_{\psi_2} = O(1)$, then with probability at least $1 - O(n^{-\Omega(1)})$ we have $\max_k |Y_k| = O(\sqrt{\log n})$ (see, e.g., Theorem 1.14 in [76]). Thus, we can choose $\gamma = \text{Poly}(\log n)$[1] to guarantee $\gamma > \max_k |Y_k|$ with high probability, and we proceed the analysis under this event. Define $\epsilon_{k,\text{appr}} := Y_{k,\text{appr}} - Y_k$, equivalently we can write

$$Y_{k,\text{appr}} = Y_k + \epsilon_{k,\text{appr}} = \big\langle \boldsymbol{X_k}, \boldsymbol{\Theta}^* \big\rangle + \epsilon_k + \epsilon_{k,\text{appr}}. \tag{1.5}$$

For $\epsilon_{k,\text{appr}}$, Lemma 1 gives $\mathbb{E}_{\Lambda_{kj}}(\gamma \cdot \dot{Y}_{kj}) = Y_k$ and hence $\mathbb{E}\epsilon_{k,\text{appr}} = 0$. Moreover, conditioned on Y_k, $\epsilon_{k,\text{appr}}$ is the mean of $f(n)$ zero-mean, independent random variables lying in $[-\gamma - Y_k, \gamma - Y_k]$. Thus, Proposition 2 and Hoeffding's Lemma (see Lemma 1.8, [76]) give

$$\|\epsilon_{k,\text{appr}}\|_{\psi_2} = O\Big(\frac{\gamma}{\sqrt{f(n)}}\Big). \tag{1.6}$$

Therefore, $\|\epsilon_{k,\text{appr}}\|_{\psi_2} = O\big(1\big)$ as long as $f(n)$ dominates γ^2, while $f(n) = \text{Poly}(\log n)$ would suffice due to $\gamma = \text{Poly}(\log n)$. In conclusion, $\mathcal{D}_{\text{mult}}$ containing $n \cdot \text{Poly}(\log n)$ binary data can

[1]Here $\text{Poly}(\log n)$ denotes any term T satisfying $T = O([\log n]^m)$ for some positive integer m.

generate the sample $\mathcal{D}_{\text{appr}}$ of size n, where each $Y_{k,\text{appr}}$ can be viewed as a full observation from (1.5). Moreover, since $\|\epsilon_k\|_{\psi_2} = \|\epsilon_{k,\text{appr}}\|_{\psi_2} = O(1)$, (1.5) is nearly equivalent to the original model (4.1). This reveals $\mathcal{D}_{\text{appr}}$, and hence $\mathcal{D}_{\text{appr}}$ with $n \cdot \text{Poly}(\log n)$ binary data, are comparable to $\mathcal{D}_{\text{full}}$ with n full data. Furthermore, this indicates the inessential logarithmic degradation of recovery error after one-bit quantization.

Note that similar heuristics can be found in sub-Gaussian regime of (sparse) covariance matrix estimation and sparse linear regression. Of course, such multi-bits heuristic deviates from the one-bit setting where we collect only one bit from each Y_k (see the following graphical illustration). But since $f(n) := \text{Poly}(\log n)$ is negligible compared with n, one may tend to believe $\mathcal{D}_{\text{mult}}$ and $\mathcal{D}_{\text{1bit}} = \{\dot{Y}_k := \text{sign}(Y_k + \Lambda_k) : k \in [n \cdot f(n)]\}$ are comparable. From this perspective, the near-optimal rates in sub-Gaussian regime are matter of courses.

$$
\begin{tikzcd}
 & \dot{Y}_{11} \arrow[rdd] & \\
 & \dot{Y}_{12} \arrow[rd] & \\
Y_1 \arrow[ruu] \arrow[ru] \arrow[rd] & \vdots & Y_{1,\text{appr}} \\
 & \dot{Y}_{1f(n)} \arrow[ru] & \\
\vdots & \dot{Y}_{n1} \arrow[rdd] & \vdots \\
 & \dot{Y}_{n2} \arrow[rd] & \\
Y_n \arrow[ruu] \arrow[ru] \arrow[rd] & \vdots & Y_{n,\text{appr}} \\
 & \dot{Y}_{nf(n)} \arrow[ru] & \\
(\mathcal{D}_{\text{full}}) & (\mathcal{D}_{\text{mult}}) & (\mathcal{D}_{\text{appr}})
\end{tikzcd}
\qquad
\begin{tikzcd}
Y_1 \arrow[r] & \dot{Y}_1 \\
Y_2 \arrow[r] & \dot{Y}_2 \\
\vdots & \vdots \\
Y_{f(n)} \arrow[r] & \dot{Y}_{f(n)} \\
Y_{(n-1)f(n)+1} \arrow[r] & \dot{Y}_{(n-1)f(n)+1} \\
Y_{(n-1)f(n)+2} \arrow[r] & \dot{Y}_{(n-1)f(n)+2} \\
\vdots & \vdots \\
Y_{nf(n)} \arrow[r] & \dot{Y}_{nf(n)} \\
 & (\mathcal{D}_{\text{1bit}})
\end{tikzcd}
$$

(A heuristic multi-bits setting) (The one-bit setting)

However, in heavy-tailed regime the story becomes totally different. Specifically, $\gamma = \text{Poly}(\log n)$ will no longer guarantee $\gamma > \max_k |Y_k|$ with high probability. When this vital condition fails, the dithering becomes invalid for responses with absolute value larger than γ. Indeed, for these measurements the proposed dithered quantization reduces to a direct collection of the sign, while under such direct quantization we even lose the well-posedness of the problem (e.g., matrix completion, see [32]) or the possibility of full signal reconstruction (e.g., one-bit compressed sensing, see [70]).

To resolve the issue, we truncate the heavy-tailed data according to some threshold η, which produces data bounded by η. Then we can treat the truncated data as sub-Gaussian data and use dithering noise drawn from $\text{uni}\big([-\gamma, \gamma]\big)$ with $\gamma > \eta$. It is not hard to see that η represents the data bias introduced in truncation, more precisely, smaller η corresponds to larger bias. Enlightened by (1.6) and Hoeffding's Lemma, γ is positively related to data variance. Definitely, for estimation or signal recovery we prefer data with small bias (i.e., big η) and small variance (i.e., small γ). But, note that we also need $\gamma > \eta$ to enforce the effectiveness of dithering. Thus, a trade-off between bias and variance is needed, and making an optimal balance between bias and variance leads to our error rates in heavy-tailed regime. See Example 1 and Example 2 in Section 5.

2 Sparse Covariance Matrix Estimation

We start from the problem of estimating a sparse covariance matrix. Let $X \in \mathbb{R}^d$ be a random vector with zero mean, the i.i.d. realizations X_k are quantized to one-bit data $(\dot{X}_{k1}, \dot{X}_{k2})$, and we aim to estimate the underlying covariance matrix $\mathbf{\Sigma}^* = \mathbb{E}XX^T$ based on the quantized data.

We first ideally assume the underlying d-dimensional random vector X has entries bounded by γ, then Corollary 1 delivers that $\mathbb{E}\big[\gamma^2 \cdot \dot{X}_{k1}\dot{X}_{k2}^T\big] = \mathbb{E}XX^T = \Sigma^*$ is just the desired covariance matrix. Besides, the concentration of $\gamma^2 \cdot \dot{X}_{k1}\dot{X}_{k2}^T$ should be fast due to boundedness, see Hoeffding's inequality in Proposition 3. Combining these observations, [36] proposed a covariance matrix estimator as an empirical version of $\mathbb{E}\big[\gamma^2 \cdot \dot{X}_{k1}\dot{X}_{k2}^T\big]$, followed by symmetrization:

$$\breve{\mathbf{\Sigma}} = \frac{\gamma^2}{2n}\sum_{k=1}^{n}\Big[\dot{X}_{k1}\dot{X}_{k2}^T + \dot{X}_{k2}\dot{X}_{k1}^T\Big]. \tag{2.1}$$

For sub-Gaussian X_k, this estimator achieves a near minimax rate (compared with full data setting in [19])

$$\|\breve{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\mathrm{op}} \lesssim \log n\sqrt{\frac{d\log d}{n}}. \tag{2.2}$$

Here, we point out that sampling two bits (rather than one bit) per entry is merely for estimating the diagonal entries of $\mathbf{\Sigma}^*$, since the one-bit version of (2.1),

$$\breve{\mathbf{\Sigma}}_{1\text{bit}} = \frac{\gamma^2}{n}\sum_{k=1}^{n}\dot{X}_{k1}\dot{X}_{k1}^T,$$

always gives γ^2 in the diagonal and hence fails to recover the diagonal of the covariance matrix.

It is evident that (2.2) requires at least $n \gtrsim d$ to provide a non-trivial error bound. Actually it has been reported that even the sample covariance matrix $\sum_k X_kX_k^T/n$ has extremely poor performance under high dimensional scaling where $d \geq n$ [51], not to mention (2.1). On the other hand, high-dimensional databases are undoubtedly becoming ubiquitous in genomics, biomedical, imaging, tomography, finance and so forth, while covariance matrix plays a fundamental role in analysis of these databases.

To address the high-dimensional issue, extra structures are necessary to reduce the intrinsic problem dimensionality. For covariance matrix we usually have sparsity as prior knowledge, especially in the situation where dependencies among different features are weak, for instance, the Genomics data [39], functional data drawn from underlying curves [74]. A precise formulation of the sparse prior is provided in Assumption 1.

Assumption 1. (Approximate column-wise sparsity) *For a specific $0 \leq q < 1$, the columns of covariance matrix $\mathbf{\Sigma}^* = [\sigma_{ij}^*]$ are approximately sparse in the sense that*

$$\sup_{j\in[d]}\sum_{i=1}^{d}|\sigma_{ij}^*|^q \leq s \tag{2.3}$$

In literature there are two mainstreams to incorporate sparsity into covariance matrix estimation, namely penalized likelihood method [12, 78] and a thresholding method [11, 18, 20, 22, 40]. Thresholding method refers to the direct regularizer that element-wisely hard thresholding the sample covariance matrix, i.e., $\mathcal{T}_\zeta(\sum_{k=1}^n X_k X_k^T/n)$, which promotes sparsity intuitively. With suitable threshold ζ, Cai and Zhou [21] showed $\mathcal{T}_\zeta(\sum_{k=1}^n X_k X_k^T/n)$ could achieve minimax rate under operator norm over the class of column-wisely sparse covariance matrices (Assumption 1). Motivated by previous work, we propose to hard thresholding $\breve{\mathbf{\Sigma}}$ in (2.1) to obtain a high-dimensional estimator $\widehat{\mathbf{\Sigma}} = [\widehat{\sigma}_{ij}]$ given by

$$\widehat{\mathbf{\Sigma}} = \mathcal{T}_\zeta \breve{\mathbf{\Sigma}}. \tag{2.4}$$

The statistical rates of $\widehat{\mathbf{\Sigma}}$ under both max norm and operator norm are established in what follows.

2.1 Sub-Gaussian Data

Assume $X_k = [X_{k,1}, X_{k,2}, ..., X_{k,d}]$ are i.i.d. sampled from a random vector $X \in \mathbb{R}^d$ with zero-mean sub-Gaussian components. In particular, we assume

$$\mathbb{E}X_k = 0,\ \|X_{k,i}\|_{\psi_2} \le \sigma,\ \forall i \in [d]. \tag{2.5}$$

From (2.4), $\breve{\mathbf{\Sigma}} = [\breve{\sigma}_{ij}]$ serves as an intermediate estimator to construct $\widehat{\mathbf{\Sigma}}$, hence we first provide an element-wise error bound of $\breve{\mathbf{\Sigma}}$ in Theorem 1.

Theorem 1. *Assume (2.5) holds. For specific $\delta \ge 1$ we assume $n > 2\delta \log d$. For some sufficiently large constant C_1 we set the dithering scale γ as*

$$\gamma = C_1 \sigma \sqrt{\log\Big(\frac{n}{2\delta \log d}\Big)} \tag{2.6}$$

and assume $\gamma > \sigma$. Then for $\breve{\mathbf{\Sigma}} = [\breve{\sigma}_{ij}]$ we have

$$\mathbb{P}\Big(|\breve{\sigma}_{ij} - \sigma^*_{ij}| \lesssim \sigma^2 \log n \sqrt{\frac{\delta \log d}{n}}\Big) \ge 1 - 2d^{-\delta} \tag{2.7}$$

for $i, j \in [d]$. Moreover, we have the error bound for max norm

$$\mathbb{P}\Big(\|\breve{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\max} \lesssim \sigma^2 \log n \sqrt{\frac{\delta \log d}{n}}\Big) \ge 1 - 2d^{2-\delta}. \tag{2.8}$$

Recall that our estimator is obtained by hard thresholding $\breve{\mathbf{\Sigma}}$. The next Theorem shows that with suitable threshold ζ, the hard thresholding even brings a tighter statistical bound for element-wise error.

Theorem 2. *Assume (2.5) holds, $\delta \ge 1$ is the same as Theorem 1, and the dithering scale γ is given as (2.6) with some C_1. Then we choose the threshold ζ by*

$$\zeta = C_2 \sigma^2 \log n \sqrt{\frac{\delta \log d}{n}}, \tag{2.9}$$

where C_2 is a sufficiently large constant. Then for any $i, j \in [d]$ we have

$$\mathbb{P}\Big(|\widehat{\sigma}_{ij} - \sigma^*_{ij}| \lesssim \min\Big\{|\sigma^*_{ij}|, \sigma^2 \log n \sqrt{\frac{\delta \log d}{n}}\Big\}\Big) \ge 1 - 2d^{-\delta}. \tag{2.10}$$

By combining (2.10) and Assumption 1, we are in a position to establish the rate of $\widehat{\mathbf{\Sigma}}$ under operator norm. Specifically, we prove that our one-bit estimator achieves a rate $O\big(s((\log n)^2\frac{\log d}{n})^{(1-q)/2}\big)$, which almost matches the minimax rate $O\big(s\left(\frac{\log d}{n}\right)^{(1-q)/2}\big)$ proved in Theorem 2 of [21]. Note that the estimator based on full data in [21] achieves the minimax rate. From this perspective, the one-bit quantization only introduces minor information loss to the learning process, i.e., a logarithmic factor. Thus, by using our method, one can embrace the privileges of one-bit data and covariance matrix of comparable accuracy simultaneously.

Theorem 3. *Assume Assumption 1, (2.5) hold, δ is the same as Theorem 1, 2 (set $\delta \geq 4$), and the dithering scale γ, the threshold ζ are respectively given by (2.6), (2.9) with some C_1, C_2. Besides, assume $\delta\log d(\log n)^2/n$ is sufficiently small. Let $p = \delta/4$, we have*

$$\Big(\mathbb{E}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\mathrm{op}}^p\Big)^{1/p} \lesssim s\Big(\sigma^2\log n\sqrt{\frac{\delta\log d}{n}}\Big)^{1-q}. \tag{2.11}$$

Moreover, the probability tail of operator norm error is bounded as

$$\mathbb{P}\Big(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\mathrm{op}} \lesssim s\Big[\sigma^2\log n\sqrt{\frac{\delta\log d}{n}}\Big]^{1-q}\Big) \geq 1 - \exp(-\delta). \tag{2.12}$$

Remark 1. *We point out that the proof of Theorem 3 may be of independent technical interest, especially the probabilistic inequality (2.12) that seems quite new in the literature. In fact, only the upper bound for the second moment (i.e., p=2 and $\delta = 8$ in (2.11)) is obtained in literature (e.g., Theorem 3 in [21]), and by Markov inequality this can only give a probability term $1-\frac{1}{\delta^{1-q}}$ in (2.12). Here, by contrast, we derive a much better probabilistic term $1-\exp(-\delta)$. The key idea is to adaptively bound the $\Omega(\delta)$-th moment rather than a specific second moment, which gives (2.11). It is straightfoward to apply this method to the traditional full-data thrsholding estimator and gain some improvements of existing results.*

To guarantee positive semi-definiteness, we introduce a trick developed in literature. Assume the eigenvalue decomposition of $\widehat{\mathbf{\Sigma}}$ is $\sum_{i=1}^d \lambda_i(\widehat{\mathbf{\Sigma}})v_iv_i^T$, we remove the components corresponding to negative eigenvalues and obtain the positive part $\widehat{\mathbf{\Sigma}}^+ = \sum_{i=1}^d \max(\lambda_i(\widehat{\mathbf{\Sigma}}), 0)v_iv_i^T$. It is not hard to show that $\|\widehat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^*\|_{\mathrm{op}} \leq 2\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\mathrm{op}}$. Thus, $\widehat{\mathbf{\Sigma}}^+$ retains the operator norm rate of $\widehat{\mathbf{\Sigma}}$. However, removing the negative components may destroy the element-wise error or the sparse pattern of $\widehat{\mathbf{\Sigma}}$, see [73].

Besides, it is worthy noting that we present Theorem 3 under operator norm by convention, but both (2.11) and (2.12) are applicable to the larger norm $\|\boldsymbol{X}\|_{1,\infty} = \sup_j\sum_i |x_{ij}|$, see an initial step in the proof (A.6).

2.2 Heavy-tailed Data

Let $X_k = [X_{k,1}, ..., X_{k,d}]^T$ be i.i.d. drawn from the random vector $X \in \mathbb{R}^d$, in this part we consider zero-mean, heavy-tailed X assumed to have bounded 4-th moment

$$\mathbb{E}X_k = 0, \mathbb{E}|X_{k,i}|^4 \leq M, \quad \forall\ i \in [d] \tag{2.13}$$

Note that this offers great relaxation compared with sub-Gaussian random variable (e.g., Proposition 1(b)) and encompasses more distributions such as t-distribution, log-normal distribution.

Compared with the light tail in Proposition 1(a), X satisfying (2.13) can have a much heavier tail, and so overlarge data appear more frequently. To illustrate why this is problematic in our quantization method, we mention that our dithering noise has finite scale γ, hence the dithering is invalid for data with magnitude larger than γ. More precisely, we have

$$\operatorname{sign}(X_{k,i} + \Gamma_{kj,i}) = \operatorname{sign}(X_{k,i}), \quad \text{if } |X_{k,i}| > \gamma.$$

Therefore, for those entries larger than γ, our signal processing reduces to a direct quantization without dithering noise, which is known to introduce great loss of information.

To deal with the issue, we first truncate the data larger than a specified threshold η and obtain the truncated data $\widetilde{X}_k$ bounded by η, which is of the spirit to introduce some biases for variance reduction. Now that the truncated data are bounded, we similarly dither them by uniform noise, and then quantize to $\dot{X}_{kj} = \operatorname{sign}(\widetilde{X}_{kj}+\Gamma_{kj})$, $j = 1, 2$, where $\Gamma_{kj} \sim \operatorname{uni}([-\gamma, \gamma]^d)$. Motivated by Corollary 1, we propose an intermediate estimator

$$\breve{\mathbf{\Sigma}} = \frac{\gamma^2}{2n} \sum_{k=1}^{n} \Big[\dot{X}_{k1}\dot{X}_{k2}^T + \dot{X}_{k2}\dot{X}_{k1}^T \Big], \tag{2.14}$$

which extends (2.1) to heavy-tailed data. Element-wise error for $\breve{\mathbf{\Sigma}}$ is given in Theorem 4.

Theorem 4. *Assume (2.13) holds. For some fixed $\delta \geq 1$ and C_3, C_4 ($C_4 > C_3$), we set the truncation parameter η and the dithering scale γ by*

$$\begin{cases} \eta = C_3 M^{1/4} \Big(\dfrac{n}{\delta \log d} \Big)^{1/8} \\ \gamma = C_4 M^{1/4} \Big(\dfrac{n}{\delta \log d} \Big)^{1/8} \end{cases}, \tag{2.15}$$

Then for $\breve{\mathbf{\Sigma}} = [\breve{\sigma}_{ij}]$ given in (2.14), we have

$$\mathbb{P}\Big(|\breve{\sigma}_{ij} - \sigma^*_{ij}| \lesssim \sqrt{M} \Big[\frac{\delta \log d}{n} \Big]^{1/4} \Big) \geq 1 - 2d^{\delta}. \tag{2.16}$$

Moreover, we have the error bound under max norm

$$\mathbb{P}\Big(\|\breve{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\max} \lesssim \sqrt{M} \Big[\frac{\delta \log d}{n} \Big]^{1/4} \Big) \geq 1 - 2d^{2-\delta}. \tag{2.17}$$

Parallel to the sub-Gaussian regime, we use an additional hard thresholding step to promote sparsity. That is, based on the intermediate estimator $\breve{\mathbf{\Sigma}}$ in (2.14), we choose some suitable thresholding parameter ζ and define the estimator

$$\widehat{\mathbf{\Sigma}} = \mathcal{T}_{\zeta} \breve{\mathbf{\Sigma}}. \tag{2.18}$$

We show the element-wise and operator norm statistical rates in Theorem 5, Theorem 6.

Theorem 5. *Assume (2.13) holds, δ is the same as Theorem 4, and the truncation threshold η and the dithering scale γ are set as (2.15) with some C_3, C_4. Then we set the threshold ζ in (2.18) by*

$$\zeta = C_5\sqrt{M}\Big(\frac{\delta\log d}{n}\Big)^{1/4} \tag{2.19}$$

where C_5 is a sufficiently large constant. Then for any $i,j\in[d]$ we have

$$\mathbb{P}\Big(|\widehat{\sigma}_{ij}-\sigma_{ij}^*|\lesssim\min\Big\{|\sigma_{ij}^*|,\sqrt{M}\Big[\frac{\delta\log d}{n}\Big]^{1/4}\Big\}\Big)\geq 1-2d^{-\delta}. \tag{2.20}$$

Theorem 6. *Assume Assumption 1, (2.13) hold, δ is fixed and the same as Theorem 4, 5 (set $\delta\geq 4$), the truncation threshold η, the dithering scale γ, the threshold ζ are set as (2.15), (2.19) for some specified C_3, C_4, C_5. Besides, assume that $\delta\log d/n$ is sufficiently small. Let $p=\delta/4$, then we have the bound for the moment of order p*

$$\Big(\mathbb{E}\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\|_{\mathrm{op}}^p\Big)^{1/p}\lesssim sM^{(1-q)/2}\Big[\frac{\delta\log d}{n}\Big]^{(1-q)/4}. \tag{2.21}$$

Moreover, we bound the probability tail of operator norm error

$$\mathbb{P}\Big(\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\|_{\mathrm{op}}\lesssim sM^{(1-q)/2}\Big[\frac{\delta\log d}{n}\Big]^{(1-q)/4}\Big)\geq 1-\exp(-\delta). \tag{2.22}$$

3 Sparse Linear Regression

We intend to establish our results for sparse linear regression (Section 3) and low-rank matrix completion (Section 4) under the unified framework of trace regression, which should be established first. Trace regression with $\mathbf{\Theta}^*\in\mathbb{R}^{d\times d}$ as desired signal is formulated as

$$Y_k = \big\langle \boldsymbol{X_k}, \mathbf{\Theta}^*\big\rangle + \epsilon_k, \tag{3.1}$$

where $\boldsymbol{X_k}\in\mathbb{R}^{d\times d}$ is covariate, ϵ_k is additive noise. To handle high-dimensional scaling, $\mathbf{\Theta}^*$ is assumed to be (approximately) low-rank (e.g., [43, 65, 66])

$$\sum_{k=1}^d|\sigma_k(\mathbf{\Theta}^*)|^q\leq r,\ \text{for some } 0\leq q<1, \tag{3.2}$$

where $\sigma_1(\mathbf{\Theta}^*)\geq\sigma_2(\mathbf{\Theta}^*)\geq\ldots\geq\sigma_d(\mathbf{\Theta}^*)$ are the singular values of $\mathbf{\Theta}^*$. For this low-rank trace regression problem, a standard approach to estimate or reconstruct $\mathbf{\Theta}^*$ is via the M-estimator (e.g., [67])

$$\widehat{\mathbf{\Theta}}\in\mathop{\arg\min}_{\mathbf{\Theta}\in\mathcal{S}}\ \mathcal{L}(\mathbf{\Theta})+\lambda\|\mathbf{\Theta}\|_{\mathrm{nu}}, \tag{3.3}$$

where $\mathcal{L}(\mathbf{\Theta})$ is a loss function that requires $\widehat{\mathbf{\Theta}}$ to fit the data $\{(\boldsymbol{X}_k, Y_k)\}$, $\|\mathbf{\Theta}\|_{\mathrm{nu}}$ is the penalty that promotes low-rankness. In [65] Negahban and Wainwright first established a general framework to obtain convergence rate for trace regression when $\mathcal{L}(\mathbf{\Theta})$ is a quadratic loss, and then many subsequent papers developed and extended the theoretical framework, to name a few, negative log-likelihood loss function [41], other estimation problems such as matrix completion with sparse corruption [57] and sparse high-dimensional time series [6], extension to quaternion field [28]. For data fitting term $\mathcal{L}(\mathbf{\Theta})$, a standard quadratic loss (i.e., ℓ_2 loss) based on full data is

$$\mathcal{L}(\mathbf{\Theta}) = \frac{1}{2n}\sum_{k=1}^{n}|Y_k - \langle \boldsymbol{X_k}, \mathbf{\Theta}\rangle|^2 = \frac{1}{2}\text{vec}(\mathbf{\Theta})^T\mathbf{\Sigma}_{XX}\text{vec}(\mathbf{\Theta}) - \langle \mathbf{\Sigma}_{Y\boldsymbol{X}}, \mathbf{\Theta}\rangle + \text{constant},$$

where $\mathbf{\Sigma}_{XX} = \sum_{k=1}^{n}\text{vec}(\boldsymbol{X_k})\text{vec}(\boldsymbol{X_k})^T/n$, $\mathbf{\Sigma}_{Y\boldsymbol{X}} = \sum_{k=1}^{n}Y_k\boldsymbol{X}_k/n$. However, this standard quadratic loss does not directly apply to our setting where full data are not available. In order to introduce some flexibility, we consider a generalized quadratic loss

$$\mathcal{L}(\mathbf{\Theta}) = \frac{1}{2}\text{vec}(\mathbf{\Theta})^T\boldsymbol{Q}\text{vec}(\mathbf{\Theta}) - \langle \boldsymbol{B}, \mathbf{\Theta}\rangle, \tag{3.4}$$

where $\boldsymbol{Q} \in \mathbb{R}^{d^2\times d^2}$ is symmetric, $\boldsymbol{B} \in \mathbb{R}^{d\times d}$. We present a framework for trace regression in Lemma 2. Note that Theorem 1 in [43] is only for $\boldsymbol{Q}, B$ in (3.4) being the (truncated) sample covariance, hence Lemma 2 can be viewed as its extension to more general $\boldsymbol{Q}, B$ that suffices for our needs. Besides, our version is refined to be more technically amenable since a useful relation (3.6) is established even without the restricted strong convexity (3.7). One shall see that (3.6) can significantly simplify the proofs for Theorem 9, 10, 11, 12.

Lemma 2. *Consider trace regression (3.1) with (approximate) low-rankness (3.2), the estimator is given by (3.3) where the loss function is a generalized quadratic loss (3.4). Let* $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$. *If* $\boldsymbol{Q}$ *is positive semi-definite, and* λ *satisfies*

$$\lambda \geq 2\|\text{mat}(\boldsymbol{Q}\cdot\text{vec}(\mathbf{\Theta}^*)) - \boldsymbol{B}\|_{\text{op}}, \tag{3.5}$$

then it holds that

$$\|\widehat{\mathbf{\Delta}}\|_{\text{nu}} \leq 10r^{\frac{1}{2-q}}\|\widehat{\mathbf{\Delta}}\|_{\text{F}}^{\frac{2-2q}{2-q}}. \tag{3.6}$$

Moreover, if the restricted strong convexity (RSC) holds, i.e., there exists $\kappa > 0$ *such that*

$$\text{vec}(\widehat{\mathbf{\Delta}})^T\boldsymbol{Q}\text{vec}(\widehat{\mathbf{\Delta}}) \geq \kappa\|\widehat{\mathbf{\Delta}}\|_{\text{F}}^2, \tag{3.7}$$

then we have the convergence rate for Frobenius norm and nuclear norm

$$\|\widehat{\mathbf{\Delta}}\|_{\text{F}} \leq 30\sqrt{r}\Big(\frac{\lambda}{\kappa}\Big)^{1-q/2} \text{ and } \|\widehat{\mathbf{\Delta}}\|_{\text{nu}} \leq 300r\Big(\frac{\lambda}{\kappa}\Big)^{1-q}. \tag{3.8}$$

With the preliminary of trace regression we now go into sparse linear regression

$$Y_k = X_k^T\Theta^* + \epsilon_k, \tag{3.9}$$

where $\Theta^* \in \mathbb{R}^d$ is the desired signal, X_k is the covariate (or sensing vector), ϵ_k is noise independent of X_k. In addition, Θ^* is approximately sparse.

Assumption 2. (Approximate sparsity on vector) *For a specific* $0 \leq q < 1$*, the desired signal* $\Theta^* = [\theta_1^*, ..., \theta_d^*]^T$ *satisfies*

$$\sum_{i=1}^{d} |\theta_i^*|^q \leq s. \tag{3.10}$$

It is not hard to see that (3.9), (3.10) are encompassed by (3.1), (3.2) if $\boldsymbol{X_k}, \boldsymbol{\Theta^*}$ are diagonal, i.e., $\boldsymbol{X_k} = \text{diag}(X_k)$, $\boldsymbol{\Theta^*} = \text{diag}(\Theta^*)$, so we consider analogue of (3.3) as the estimator. The first issue is the choice of loss function since the existing methods are invalid: we can neither use the quadratic loss as [43,65] without full data, nor the negative log-likelihood as [41] due to the noise ϵ_k with unknown distribution. Instead, we resort to a generalized quadratic loss given in (3.4) to proceed. For sparse linear regression, particularly, we let $\mathcal{L}(\Theta) = \frac{1}{2}\Theta^T \boldsymbol{Q}\Theta - B^T\Theta$ where $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ is symmetric, $B \in \mathbb{R}^d$. Thus, our estimator is given by

$$\widehat{\Theta} \in \mathop{\arg\min}_{\Theta\in\mathbb{R}^d} \ \frac{1}{2}\Theta^T\boldsymbol{Q}\Theta - B^T\Theta + \lambda\|\Theta\|_1. \tag{3.11}$$

Lemma 2 implies the following Corollary.

Corollary 2. *Consider linear regression (3.9) with (approximate) sparsity (3.10), the estimator* $\widehat{\Theta}$ *is given by (3.11). Let* $\widehat{\Delta} = \widehat{\Theta} - \Theta^*$*. If* $\boldsymbol{Q}$ *is positive semi-definite,* λ *satisfies*

$$\lambda \geq 2\|\boldsymbol{Q}\Theta^* - B\|_{\max}, \tag{3.12}$$

then it holds that

$$\|\widehat{\Delta}\|_1 \leq 10 s^{\frac{1}{2-q}} \|\widehat{\Delta}\|_2^{\frac{2-2q}{2-q}}. \tag{3.13}$$

Moreover, if for some $\kappa > 0$ *we have the restricted strong convexity*

$$\widehat{\Delta}^T\boldsymbol{Q}\widehat{\Delta} \geq \kappa\|\widehat{\Delta}\|_2^2, \tag{3.14}$$

then we have the error bound for ℓ_2 *and* ℓ_1 *norm*

$$\|\widehat{\Delta}\|_2 \leq 30\sqrt{s}\Big(\frac{\lambda}{\kappa}\Big)^{1-q/2} \quad \text{and} \quad \|\widehat{\Delta}\|_1 \leq 300 s\Big(\frac{\lambda}{\kappa}\Big)^{1-q} \tag{3.15}$$

It remains to properly specify $\boldsymbol{Q}, B$ in (3.11). Note that the expected quadratic risk is given by

$$\mathbb{E}|Y_k - X_k^T\Theta|^2 = \Theta^T\mathbb{E}(X_kX_k^T)\Theta - (\mathbb{E}(Y_kX_k))^T\Theta + \text{constant},$$

thus a general guideline to choose $\boldsymbol{Q}$, B is implied, that is, $\boldsymbol{Q}$ should be close to the covariance matrix of X_k, and B should well approximate the covariance $\mathbb{E}(Y_kX_k)$. Naturally, based on one-bit data we can still use $\widehat{\boldsymbol{\Sigma}}$ in (2.4) or $\breve{\boldsymbol{\Sigma}}$ in (2.1) as $\boldsymbol{Q}$. Nevertheless, the issue is that they may not be positive semi-definite, while the positive semi-definiteness of $\boldsymbol{Q}$ is an indispensable condition in Corollary 2.

To close the gap, we assume $\boldsymbol{\Sigma}_{XX} = \mathbb{E}X_kX_k^T$ is column-wisely sparse.We mention that such sparsity is quite common in high-dimensional regime, and it definitely accommodates isotropic sensing vectors that is widely adopted in compressed sensing (See Remark 2).

Assumption 3. *$X_1, ..., X_n$ are i.i.d. drawn from a zero-mean random vector with covariance matrix $\mathbf{\Sigma}_{XX} = \mathbb{E}X_kX_k^T = [\sigma_{ij}]$ satisfying Assumption 1 under parameter $(0, s_0)$, i.e., the number of non-zero elements in each column is less than s_0. Besides, $\mathbf{\Sigma}_{XX}$ is positive definite, and for some absolute constant $\kappa_0 > 0$ it satisfies $\lambda_{\min}(\mathbf{\Sigma}_{XX}) \geq 2\kappa_0$.*

Under Assumption 3, our estimator $\widehat{\mathbf{\Sigma}}$ defined in (2.4) for sub-Gaussian data, or (2.18) for heavy-tailed data, is positive definite with high probability. Thus, we set $\boldsymbol{Q} = \widehat{\mathbf{\Sigma}}$ in (3.11). Note that $\mathbb{E}(Y_kX_k)$ is also covariance, enlightened by Corollary 1, we similarly set

$$\widehat{\Sigma}_{YX} = \frac{1}{n}\sum_{k=1}^{n} \gamma^2 \cdot \dot{Y}_k\dot{X}_{k1}. \tag{3.16}$$

Now we have specified our estimator as

$$\widehat{\Theta} \in \arg\min_{\Theta\in\mathbb{R}^d} \ \frac{1}{2}\Theta^T\widehat{\mathbf{\Sigma}}\Theta - \widehat{\Sigma}_{YX}^T\Theta + \lambda\|\Theta\|_1. \tag{3.17}$$

3.1 Sub-Gaussian Data

We assume the sub-Gaussian covariate and zero-mean sub-Gaussian noise satisfy $\|X_k\|_{\psi_2} \leq \sigma_1$, $\|\epsilon_k\|_{\psi_2} \leq \sigma_2$, and $\|\Theta^*\|_2 \leq R = O(1)$. In this setting, we have $\|Y_k\|_{\psi_2} \leq \|X_k^T\Theta^*\|_{\psi_2} + \|\epsilon_k\|_{\psi_2} \leq \|\Theta^*\|_2\|X_k\|_{\psi_2} + \|\epsilon_k\|_{\psi_2} = O(\max\{\sigma_1, \sigma_2\})$. To lighten notations without losing generality, we assume for some $\sigma > 0$

$$\max\left\{\|X_k\|_{\psi_2}, \|Y_k\|_{\psi_2}\right\} \leq \sigma \tag{3.18}$$

and use the uniform noise with the same dithering scale γ to dither X_k and Y_k before one-bit quantization. More precisely, we choose dithering noise $\Gamma_{k1}, \Gamma_{k2} \sim \text{uni}([-\gamma,\gamma]^d), \Lambda_k \sim \text{uni}([-\gamma,\gamma])$ with γ in (2.6), then we obtain the one-bit data $(\dot{X}_{k1}, \dot{X}_{k2}, \dot{Y}_k)$.

We mention that our result directly extends to more general setting where $\|X_k\|_{\psi_2}, \|Y_k\|_{\psi_2}$ may vary a lot. Indeed, we can adaptively choose dithering scale according to $\|X_k\|_{\psi_2}$ and $\|Y_k\|_{\psi_2}$, for instance, $\Gamma_{k1}, \Gamma_{k2} \sim \text{uni}([-\gamma_X, \gamma_X]^d)$, $\Lambda_k \sim \text{uni}([\gamma_Y, \gamma_Y])$. In our numerical simulations, we also applied different dithering scales to X_k and Y_k to improve the recovery.

In Theorem 7 we will give the near minimax statistical rate for the estimator $\widehat{\Theta}$. The idea is to invoke Corollary 2, and this requires (3.12) and (3.14). To properly set λ to confirm (3.12), it suffices to bound $\|\widehat{\mathbf{\Sigma}}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max}$ from above. Combining Assumption 3 and results in Section 2, we can show (3.14) holds with high probability.

Theorem 7. *Assume (3.9), Assumption 2, (3.18) hold, $\|\Theta^*\|_2 \leq R$ for some absolute constant R, and we have Assumption 3 for the covariate X_k. Before the quantization we dither the data with γ in (2.6). We consider $\widehat{\Theta}$ given by (3.17) where $\widehat{\mathbf{\Sigma}}, \widehat{\Sigma}_{YX}$ are respectively set as (2.4), (3.16), and ζ is given by (2.9). Moreover, we choose λ by*

$$\lambda = C_6 \log n\sqrt{\frac{\delta\log d}{n}} \tag{3.19}$$

with sufficiently large C_6. Let $\widehat{\Delta} = \widehat{\Theta} - \Theta^*$. When $(\log n)^2\log d/n$ is sufficiently small, with probability at least $1 - \exp(-\delta) - 2d^{2-\delta}$, we have

$$\begin{cases} \|\widehat{\Delta}\|_2 \lesssim \sqrt{s}\Big(\sigma^2\log n\sqrt{\dfrac{\delta\log d}{n}}\Big)^{1-q/2} \\ \|\widehat{\Delta}\|_1 \lesssim \ s\Big(\sigma^2\log n\sqrt{\dfrac{\delta\log d}{n}}\Big)^{1-q} \end{cases}. \tag{3.20}$$

Remark 2. *Compared with the sample covariance* $\sum_{k=1}^{n} X_k^T X_k/n$*, the proposed one-bit covariance matrix estimator* $\widehat{\mathbf{\Sigma}}$ *lacks positive semi-definiteness. We address the issue by assuming column-wise sparsity of* $\mathbf{\Sigma}_{XX}$*, which together with* $\lambda_{\min}(\mathbf{\Sigma}_{XX}) = \Omega(1)$ *can provide positive definiteness under high-dimensional scaling. This assumption is also used in [91] to resolve the same issue. As an example, this accommodates isotropic sensing vectors that is conventionally adopted in compressed sensing literature, see [25, 38, 72] for instance.*

3.2 Heavy-tailed Data

We then switch to the heavy-tailed case where X_k and ϵ_k are only assumed to possess bounded 4-th moment. We consider the scaling of the desired signal as $\|\Theta^*\|_2 \le R = O(1)$. Moreover, we assume $\mathbb{E}|V^T X_k|^4 \le M_1$ for any $V \in \mathbb{R}^d$, $\|V\|_2 \le 1$, and $\mathbb{E}|\epsilon_k|^4 \le M_2$. Then we have the fourth moment of Y_k is also bounded by $O\big(R^4 M_1 + M_2\big)$. To lighten the notation without losing generality, we assume the same upper bound M for covariate and response:

$$\max\Big\{\sup_{\|V\|_2\le 1} \mathbb{E}|V^T X_k|^4, \mathbb{E}|Y_k|^4\Big\} \le M, \tag{3.21}$$

which allows us to use the same truncation parameter η and dithering scale γ for X_k and Y_k.

Similar to the same comment for sub-Gaussian case, if the fourth moment of X_k and Y_k have different scales, our method still works under different truncation parameters (η_X, η_Y) and dithering parameters (γ_X, γ_Y). Moreover, it is straightforward to adapt our method to the mixing case studied in [43] where X_k is sub-Gaussian but ϵ_k (and hence Y_k) is heavy-tailed. In this mixing setting, only the responses are treated as heavy-tailed data and truncated before dithering.

Theorem 8. *Assume (3.9), Assumption 2, (3.21) hold,* $\|\Theta^*\|_2 \le R$ *for some absolute constant* R*, and we have Assumption 3 for the covariate* X_k*. By setting* η, γ *as (2.15) such that* $\gamma > \eta$*, we first truncate* (X_k, Y_k) *element-wisely to* $(\widetilde{X}_k, \widetilde{Y}_k)$ *with parameter* η*, then dither the truncated data with uniform noise on* $[-\gamma, \gamma]$ *and quantize the data to* $(\dot{X}_{k1}, \dot{X}_{k2}, \dot{Y}_k)$ *finally. We consider* $\widehat{\Theta}$ *in (3.17) where* $\widehat{\mathbf{\Sigma}}$ *is given in (2.18) with* ζ *set as (2.19),* $\widehat{\Sigma}_{YX}$ *is given in (3.16). Moreover, we choose*

$$\lambda = C_7\sqrt{M}\Big(\frac{\delta \log d}{n}\Big)^{1/4} \tag{3.22}$$

with sufficiently large C_7*. Let* $\widehat{\Delta} = \widehat{\Theta} - \Theta^*$*. When* $\log d/n$ *is sufficiently small, with probability at least* $1 - \exp(-\delta) - 2d^{2-d}$*, we have*

$$\begin{cases} \|\widehat{\Delta}\|_2 \lesssim \sqrt{s} M^{1/2-q/4}\Big(\dfrac{\delta \log d}{n}\Big)^{1/4-q/8} \\ \|\widehat{\Delta}\|_1 \lesssim \ sM^{(1-q)/2}\Big(\dfrac{\delta \log d}{n}\Big)^{(1-q)/4} \end{cases}. \tag{3.23}$$

We emphasize that our method does not rely on the full knowledge of $\mathbf{\Sigma}_{XX}$, instead, our method applies as long as $\mathbf{\Sigma}_{XX}$ satisfies Assumption 3. In a variant setting where $\mathbf{\Sigma}_{XX}$ is known as a priori, we can directly set $\boldsymbol{Q} = \mathbf{\Sigma}_{XX}$ in (3.11), and the same error rates can be obtained by the same techniques. Instead of giving the details, we only note two differences: Firstly, the sparsity in Assumption 3 can be removed (as we do not need the estimator $\widehat{\mathbf{\Sigma}}$); Secondly, only one bit per entry for X_k is sufficient since $\widehat{\Sigma}_{YX}$ does not involve $\dot{X}_{k2}$, see (3.16).

3.3 One-bit Compressed Sensing

We just studied sparse linear regression based on the one-bit quantized covariates and responses $(\dot{X}_{k1}, \dot{X}_{k2}, \dot{Y}_k)$, while the only related problem studied in existing works is one-bit compressed sensing (1-bit CS). In 1-bit CS, one considers the same linear model (3.9) and wants to estimate the sparse underlying signal Θ^* based on $(X_k, \dot{Y}_k)$, where X_k denotes the full covariate, and $\dot{Y}_k \in \{-1, 1\}$ is the one-bit quantized version of the response Y_k. In particular, earlier works mainly studied a direct quantization with $\dot{Y}_k = \text{sign}(X^T\Theta^*)$ (see, e.g., [15, 50, 69, 70]), while recent works (e.g., [5, 37, 38, 58, 83]) began to consider dithered quantization which are more related to our work, i.e., $\dot{Y}_k = \text{sign}(X^T\Theta^* + \Lambda)$ for some dithering noise Λ. Specifically, by the additional dithering step, these works overcome several limitations and present better results. For instance, full reconstruction with norm [58], exponentially-decaying error rate [5], and extension to non-Gaussian sensing vectors [37, 38, 83].

Since one still has full knowledge on X_k in 1-bit CS, our problem setting is novel and evidently more tricky. From a practical viewpoint, due to the binary covariate, the storage and communication costs are further lowered in our method. Technically, the key element that allows quantization of covariate is the new 1-bit sparse covariance matrix estimator developed in Section 2. To facilitate presentation and future study, we term this new setting as *one-bit quantized-covariate compressed sensing (1-bit QC-CS)* to distinguish with the canonical 1-bit CS. Thus, it is unfair to compare our Theorem 7, 8 with existing results for 1-bit CS.

To see the contributions of this paper more explicitly, we analogously establish results for 1-bit CS where full-precision X_k are available, under both sub-Gaussian and heavy-tailed regimes. Similar to (3.17), we formulate the estimation as a convex programming problem, but substitute $\widehat{\mathbf{\Sigma}}$ in (3.17) by the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{XX} = \sum_{k=1}^n X_k X_k^T/n$ for sub-Gaussian X_k, or the truncated sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} = \sum_{k=1}^n \widetilde{X}_k \widetilde{X}_k^T/n$ for heavy-tailed X_k. Here, for heavy-tailed case, we truncate X_k element-wisely, but we distinguish the truncation threshold of X_k, Y_k by different notations η_X, η_Y and they are set to be different values. More precisely, the i-th entry of $\widetilde{X}_k$ is given by $\widetilde{X}_{k,i} = \text{sign}(X_{k,i}) \min\{|X_{k,i}|, \eta_X\}$, while before the dithered quantization, Y_k is truncated to be $\widetilde{Y}_k = \text{sign}(Y_k) \min\{|Y_k|, \eta_Y\}$.

Although the results are similarly established by the framework of trace regression, we feel obliged to note some differences. Let us consider the sub-Gaussian regime for illustration. Firstly, the column-wise sparsity of $\mathbf{\Sigma}_{XX}$ in Assumption 3, whose main aim is to guarantee positive semi-definiteness of the one-bit covariance matrix estimator $\widehat{\mathbf{\Sigma}}$, can now be removed as $\widehat{\mathbf{\Sigma}}_{XX}$ is automatically positive semi-definite. But on the other hand, without this assumption, we no longer have a dimension-free upper bound on $\|\widehat{\mathbf{\Sigma}}_{XX} - \mathbf{\Sigma}_{XX}\|_{\text{op}}$, hence the proof cannot proceed to (B.11). Indeed, one only has dimension-free upper bound on $\|\widehat{\mathbf{\Sigma}}_{XX} - \mathbf{\Sigma}_{XX}\|_{\max}$, and we need to impose a stronger scaling $\|\Theta^*\|_1 = O(1)$ to close the gap (This is also assumed in heavy-tailed case of sparse linear regression in [43], see Lemma 1(b) therein). Beyond that, (B.9) is also missing, so we need to establish the restricted strong convexity (3.7) in Lemma 2 via some additional technicalities.

In the next two theorems we present our results on 1-bit CS, which are directly comparable to the vast literature of 1-bit CS. To facilitate the flow of our presentation, a detailed comparison is postponed to Appendix D. One shall see that, even without mentioning our first study of 1-bit QC-CS, the following two results concerning 1-bit CS already gain significant improvements upon existing works.

Theorem 9. (1-bit CS with sub-Gaussian data) *Assume (3.9), Assumption 2, (3.18) hold (σ*

in (3.18) is an absolute constant), $\|\Theta^*\|_1 \leq R$ *for some absolute constant* R*. For the zero-mean covariate* X_k*, define the covariance matrix* $\mathbf{\Sigma}_{XX} = \mathbb{E}X_kX_k^T$ *and we assume* $\lambda_{\min}(\mathbf{\Sigma}_{XX}) \geq 2\kappa_0$ *for some absolute constant* $\kappa_0 > 0$*. We quantize* Y_k *to be* $\dot{Y}_k = \mathrm{sign}(Y_k + \Lambda_k)$ *with* Λ_k *uniformly distributed on* $[-\gamma, \gamma]$*, and* γ *is set as (2.6) for specific* $\delta > 0$*. The estimation is formulated as a convex programming problem*

$$\widehat{\Theta} \in \arg\min_{\Theta \in \mathbb{R}^d} \ \frac{1}{2}\Theta^T\widehat{\mathbf{\Sigma}}_{XX}\Theta - \widehat{\Sigma}_{YX}^T\Theta + \lambda\|\Theta\|_1. \tag{3.24}$$

Moreover, we set $\widehat{\mathbf{\Sigma}}_{XX}$*,* $\widehat{\Sigma}_{YX}^T$ *and* λ *in (3.24) as*

$$\widehat{\mathbf{\Sigma}}_{XX} = \frac{1}{n}\sum_{k=1}^{n} X_kX_k^T, \quad \widehat{\Sigma}_{YX} = \frac{1}{n}\sum_{k=1}^{n} \gamma \cdot \dot{Y}_K X_k, \quad \lambda = C_8\sqrt{\frac{\delta \log d \log n}{n}} \tag{3.25}$$

with some sufficiently large C_8*. Let* $\widehat{\Delta} = \widehat{\Theta} - \Theta^*$*, then when* $s\big(\frac{\delta \log d}{n}\big)^{1-q/2}$ *is sufficiently small, with probability at least* $1 - 7d^{2-\delta}$*, we have*

$$\begin{cases} \|\widehat{\Delta}\|_2 \lesssim \sqrt{s}\Big(\sqrt{\dfrac{\delta \log d \log n}{n}}\Big)^{1-q/2} \\ \|\widehat{\Delta}\|_1 \lesssim \ s\Big(\sqrt{\dfrac{\delta \log d \log n}{n}}\Big)^{1-q} \end{cases}. \tag{3.26}$$

By taking advantage of the full covariate and new technical tools, in heavy-tailed regime of 1-bit CS we show the ℓ_2 norm error rate $O\big(s^{2/3}\big(\frac{\log d}{n}\big)^{1/3-q/6}\big)$, which is faster than the corresponding rate for 1-bit QC-CS in Theorem 8.

Theorem 10. (1-bit CS with heavy-tailed data) *Assume (3.9), Assumption 2, (3.21) hold (M in (3.21) is an absolute constant),* $\|\Theta^*\|_1 \leq R$ *for some absolute constant* R*. For the zero-mean covariate* X_k *we let* $\mathbf{\Sigma}_{XX} = \mathbb{E}X_kX_k^T$ *and assume* $\lambda_{\min}(\mathbf{\Sigma}_{XX}) \geq 2\kappa_0$ *for some absolute constant* $\kappa_0 > 0$*. We element-wisely truncate* X_k *to be* $\widetilde{X}_k$ *with threshold* η_X*, and truncate* Y_k *to be* $\widetilde{Y}_k$ *with threshold* η_Y*. Then,* $\widetilde{Y}_k$ *is dithered and quantized to be* $\dot{Y}_k = \mathrm{sign}(\widetilde{Y}_k + \Lambda_k)$ *with* Λ_k *uniformly distributed on* $[-\gamma, \gamma]$*. For specific* $\delta > 0$*, we set these signal processing parameters as*

$$\eta_X = C_9\Big(\frac{n}{\delta \log d}\Big)^{\frac{1}{4}}, \quad \eta_Y = C_{10}\Big(\frac{n}{\delta \log d}\Big)^{\frac{1}{6}}, \quad \gamma = C_{11}\Big(\frac{n}{\delta \log d}\Big)^{\frac{1}{6}}, \tag{3.27}$$

where $C_{11} > C_{10}$ *to give* $\gamma > \eta_Y$*. The estimation is formulated as a convex programming problem*

$$\widehat{\Theta} \in \arg\min_{\Theta \in \mathbb{R}^d} \ \frac{1}{2}\Theta^T\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}\Theta - \widehat{\Sigma}_{YX}^T\Theta + \lambda\|\Theta\|_1. \tag{3.28}$$

Moreover, we set $\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}$*,* $\widehat{\Sigma}_{YX}$ *and* λ *in (3.28) as*

$$\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} = \frac{1}{n}\sum_{k=1}^{n} \widetilde{X}_k\widetilde{X}_k^T, \quad \widehat{\Sigma}_{YX} = \frac{1}{n}\sum_{k=1}^{n} \gamma \cdot \dot{Y}_k\widetilde{X}_k, \quad \lambda = C_{12}\Big(\frac{\delta \log d}{n}\Big)^{1/3} \tag{3.29}$$

with some sufficiently large C_{12}. Let $\widehat{\Delta} = \widehat{\Theta} - \Theta^*$, then when $s^2\big(\frac{\delta \log d}{n}\big)^{1-q/2}$ is sufficiently small, with probability at least $1 - O(d^{2-\sqrt{\delta}})$, we have

$$\begin{cases} \|\widehat{\Delta}\|_2 \lesssim s^{\frac{2}{3}}\Big(\dfrac{\delta \log d}{n}\Big)^{\frac{1-q/2}{3}} \\ \|\widehat{\Delta}\|_1 \lesssim s^{\frac{4}{3}}\Big(\dfrac{\delta \log d}{n}\Big)^{\frac{1-q}{3}} \end{cases}. \tag{3.30}$$

To conclude this section, we briefly state how to adjust Theorem 9, 10 to the situation where $\mathbf{\Sigma}_{XX}$ is known as a priori. Likewise, we use the known $\mathbf{\Sigma}_{XX}$ to replace $\widehat{\mathbf{\Sigma}}_{XX}$ in (3.24), or $\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}$ in (3.28). Because some error terms vanish, and the restricted strong convexity holds automatically, the technical proofs of the error rates can be greatly shortened.

4 Low-rank Matrix Completion

Matrix completion refers to the problem of recovering a low-rank matrix with incomplete observations of the entries, which is motivated by recommendation system, system identification, quantum state tomography, image inpainting, and many others, see [7, 28–30, 33, 44, 47] for instance. The literature can be roughly divided into two lines, exact recovery and approximate recovery (i.e., statistical estimation). To establish exact recovery guarantee, the underlying matrix is required to satisfy a quite stringent incoherence condition proposed and developed in [24, 26, 27, 75]. By contrast, it was shown that matrix with low spikiness could be well approximated (or estimated) under much more relaxed condition [55, 59, 66]. This Section is intended to study the estimation problem of matrix completion via the binary data produced by our one-bit quantization scheme. For simplicity we consider square matrix and formulate the model as

$$Y_k = \langle \boldsymbol{X_k}, \mathbf{\Theta}^* \rangle + \epsilon_k \tag{4.1}$$

where $\mathbf{\Theta}^*$ is the underlying low-rank data matrix of interest, $\boldsymbol{X_k}$ distributed on $\{e_i e_j^T : i, j \in [d]\}$ is the sampler that extracts one entry of $\mathbf{\Theta}^*$, Y_k is the k-th observation corrupted by noise ϵ_k independent of $\boldsymbol{X_k}$. We consider a random, uniform sampling scheme

$$\boldsymbol{X_1}, ..., \boldsymbol{X_n} \text{ are i.i.d. uniformly distributed on } \{e_i e_j^T : i \in [d], j \in [d]\}, \tag{4.2}$$

but we mention that the results can be directly adapted to non-uniform sampling scheme, see [55]. To embrace more real applications, $\mathbf{\Theta}^*$ is assumed to be approximately low-rank [66].

Assumption 4. (Approximate low-rankness on matrix) *Let $\sigma_1(\mathbf{\Theta}^*) \geq ... \geq \sigma_d(\mathbf{\Theta}^*)$ be singular values of $\mathbf{\Theta}^*$, $0 \leq q < 1$. For some $r > 0$ it holds that*

$$\sum_{k=1}^{d} \sigma_k(\mathbf{\Theta}^*)^q \leq r. \tag{4.3}$$

Since $\boldsymbol{X_k}$ only has d^2 values, we can use $\lceil 2 \log_2 d \rceil$ bits to encode $\boldsymbol{X_k}$ without losing any information. Therefore, we only quantize Y_k to binary data $\dot{Y}_k$ and study the estimation via $(\boldsymbol{X_k}, \dot{Y}_k)$. Based on previous experience in 1-bit QC-CS and 1-bit CS, we use a generalized quadratic loss (3.4) with $\boldsymbol{Q},\boldsymbol{B}$ specified to be

$$\boldsymbol{Q}=\frac{1}{n}\sum_{k=1}^{n}\mathrm{vec}(\boldsymbol{X_k})\mathrm{vec}(\boldsymbol{X_k})^T\ ,\quad \boldsymbol{B}=\frac{1}{n}\sum_{k=1}^{n}\gamma\cdot\dot{Y}_k\boldsymbol{X_k}. \tag{4.4}$$

The spikiness of $\boldsymbol{\Theta^*}$ is defined to be $\frac{d\|\boldsymbol{\Theta^*}\|_{\max}}{\|\boldsymbol{\Theta^*}\|_{\mathrm{F}}}$ in [66], note that completing a matrix with high spikiness (close to d) with incomplete observations could be an ill-posed problem per se [33], hence [66] assumed $\boldsymbol{\Theta^*}$ to have bounded spikiness. Besides the spikiness, a similar but more straightforward assumption is a max-norm constraint (e.g., [23, 28, 32, 55]), which intuitively excludes the appearance of overlarge entry and hence some pathological cases such as $\boldsymbol{\Theta^*}=e_{i_0}e_{j_0}^T$ for some $(i_0,j_0)\in[d]\times[d]$. Here, we adopt this more direct condition and assume

$$\|\boldsymbol{\Theta^*}\|_{\max}\le\alpha^*. \tag{4.5}$$

Substitute (4.4), (3.4) into (3.3), together with the max-norm constraint (4.5), we now define our estimator via the following convex programming problem

$$\begin{aligned}\widehat{\boldsymbol{\Theta}}&\in\mathop{\arg\min}_{\|\boldsymbol{\Theta}\|_{\max}\le\alpha^*}\ \frac{1}{2}\mathrm{vec}(\boldsymbol{\Theta})^T\boldsymbol{Q}\mathrm{vec}(\boldsymbol{\Theta})-\langle\boldsymbol{B},\boldsymbol{\Theta}\rangle+\lambda\|\boldsymbol{\Theta}\|_{\mathrm{nu}}\\&=\mathop{\arg\min}_{\|\boldsymbol{\Theta}\|_{\max}\le\alpha^*}\ \frac{1}{2n}\sum_{k=1}^{n}\big(\big\langle\boldsymbol{X_k},\boldsymbol{\Theta}\big\rangle-\gamma\cdot\dot{Y}_k\big)^2+\lambda\|\boldsymbol{\Theta}\|_{\mathrm{nu}}\end{aligned} \tag{4.6}$$

Compared with the program (3.17) involving $\widehat{\boldsymbol{\Sigma}}$ used in sparse linear regression, (4.6) is more intuitive since we simply replace the full observation Y_k in a standard quadratic loss with its one-bit surrogate $\gamma\cdot\dot{Y}_k$. We draw this inspiration from Lemma 1.

Applying Lemma 2 to the problem set-up of low-rank matrix completion directly gives Corollary 3, hence its proof is omitted.

Corollary 3. *Consider (4.1) under random sampling (4.2), $\boldsymbol{\Theta^*}$ satisfies Assumption 4 and (4.5). Consider $\widehat{\boldsymbol{\Theta}}$ in (4.6). Let $\widehat{\boldsymbol{\Delta}}=\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta^*}$. If*

$$\lambda\ge2\Big\|\frac{1}{n}\sum_{k=1}^{n}\big[\,\langle\boldsymbol{X_k},\boldsymbol{\Theta^*}\rangle-\gamma\cdot\dot{Y}_k\big]\boldsymbol{X_k}\Big\|_{\mathrm{op}}, \tag{4.7}$$

then it holds that

$$\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}}\le10r^{\frac{1}{2-q}}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^{\frac{2-2q}{2-q}}. \tag{4.8}$$

Moreover, if the RSC holds, i.e., for some $\kappa>0$

$$\frac{1}{n}\sum_{k=1}^{n}|\big\langle\boldsymbol{X_k},\widehat{\boldsymbol{\Delta}}\big\rangle|^2\ge\kappa\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2, \tag{4.9}$$

then we have

$$\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}\le30\sqrt{r}\Big(\frac{\lambda}{\kappa}\Big)^{1-q/2}\ \text{ and }\ \|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}}\le300r\Big(\frac{\lambda}{\kappa}\Big)^{1-q}. \tag{4.10}$$

4.1 Sub-Gaussian noise

We first cope with sub-Gaussian noise ϵ_k satisfying

$$\mathbb{E}\epsilon_k = 0,\ \|\epsilon_k\|_{\psi_2} \le \sigma. \tag{4.11}$$

Recall our dithered quantization scheme is formulated as $\dot{Y}_k = \mathrm{sign}(Y_k + \Lambda_k)$ with Λ_k uniformly distributed on $[-\gamma, \gamma]$. To invoke Corollary 3 and obtain the statistical rate, we need to choose suitable λ that guarantees (4.7) with high probability, and this requires an upper bound for the right hand side of (4.7). For clarity, we present this part as the following Lemma.

Lemma 3. *Consider (4.1) under sampling scheme (4.2), max-norm constraint (4.5), and sub-Gaussian noise assumption (4.11). For a specific $\delta > 1$, we choose the dithering noise scale γ by*

$$\gamma = C_{13} \max\{\alpha^*, \sigma\} \sqrt{\log\Big(\frac{n}{\delta d \log(2d)}\Big)} \tag{4.12}$$

with some sufficiently large C_{13} such that $\gamma \ge 2\max\{\alpha^*, \sigma\}$. If $\frac{\delta d \log d}{n}$ is sufficiently small, we have

$$\Big\|\frac{1}{n}\sum_{k=1}^n \big[\langle \boldsymbol{X_k}, \boldsymbol{\Theta^*}\rangle - \gamma \cdot \dot{Y}_k\big]\boldsymbol{X_k}\Big\|_{\mathrm{op}} \lesssim \max\{\alpha^*, \sigma\}\sqrt{\log n \frac{\delta \log d}{nd}} \tag{4.13}$$

with probability higher than $1 - 2d^{1-\delta}$.

It remains to consider (4.9). To lighten the notation we use $\mathscr{X} = (\boldsymbol{X_1}, ..., \boldsymbol{X_n})$ to denote the observed positions and define $\mathcal{F}_{\mathscr{X}}(\boldsymbol{\Theta}) = n^{-1}\sum_{k=1}^n |\big\langle \boldsymbol{X_k}, \boldsymbol{\Theta}\big\rangle|^2$. It is known that $\mathcal{F}_{\mathscr{X}}(\boldsymbol{\Theta}) \ge \kappa\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2$ may not always hold under high-dimensional scaling and the special covariate (4.2). In this case, one often needs to establish (4.9) with a relaxed term (i.e., tolerance function), see Definition 2 in [67] for instance.

Based on this idea, in Theorem 1 of [66], Negahban and Wainwright first established such relaxed RSC over a constraint set. Later, in Lemma 12 of [55], Klopp considered a different constraint set and provided a refined proof, but only for the exact low-rank setting, i.e., $q = 0$ in Assumption 4. More recently, in Lemma 5 of [28], Chen and Ng considered a constraint set depending on $q \in [0, 1)$ and extended the proof in [55] to approximate low-rank regime. As a consequence, a simpler and much shorter proof for the error bound in [66] could be obtained, see more discussions in [28]. Here we show the relaxed RSC over the constraint set defined in [28], see $\mathcal{C}(\psi)$ in (4.14).

Lemma 4. *For a specific δ and sufficiently large ψ, we consider the constraint set*

$$\begin{aligned}\mathcal{C}(\psi) = \Big\{\boldsymbol{\Theta} \in \mathbb{R}^{d\times d} : &\|\boldsymbol{\Theta}\|_{\max} \le 2\alpha^*, \|\boldsymbol{\Theta}\|_{\mathrm{nu}} \le 10 r^{\frac{1}{2-q}} \|\boldsymbol{\Theta}\|_{\mathrm{F}}^{\frac{2-2q}{2-q}}, \\ &\|\boldsymbol{\Theta}\|_{\mathrm{F}}^2 \ge (\alpha^* d)^2 \sqrt{\frac{\psi\delta\log(2d)}{n}}\Big\}.\end{aligned} \tag{4.14}$$

Then there exists some absolute constant $\kappa \in (0, 1)$, such that with probability at least $1 - d^{-\delta}$, it holds that

$$\mathcal{F}_{\mathscr{X}}(\boldsymbol{\Theta}) \ge \kappa d^{-2}\|\boldsymbol{\Theta}\|_{\mathrm{F}}^2 - T_0,\ \forall \boldsymbol{\Theta} \in \mathcal{C}(\psi), \tag{4.15}$$

where the relaxation term T_0 is given by

$$T_0 = \frac{r}{(2-q)d^q}\Big(240\alpha^*\sqrt{\frac{d\log(2d)}{n}}\Big)^{2-q}. \tag{4.16}$$

We are now ready to derive the statistical bound of the estimation error $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$. The main idea is parallel to previous works [28,55,66], i.e., to discuss whether $\widehat{\mathbf{\Delta}}$ belongs to $\mathcal{C}(\psi)$. Note that this only hinges on the third constraint in (4.14), since the first two constraints are automatically satisfied by $\widehat{\mathbf{\Delta}}$, see (4.5) and (4.8).

Theorem 11. *Under the setting of Lemma 3, assume $\mathbf{\Theta}^*$ satisfies Assumption 4, we consider the estimator $\widehat{\mathbf{\Theta}}$ defined in (4.6). Moreover, we set λ by*

$$\lambda = C_{14}\max\{\alpha^*,\sigma\}\sqrt{\log n\frac{\delta\log d}{nd}} \tag{4.17}$$

with sufficiently large C_{14}, assume $\frac{\delta d\log d}{n}$ is sufficiently small, $r \gtrsim d^q$, $n \lesssim d^2\log(2d)$, then with probability higher than $1-3d^{1-\delta}$, we have

$$\begin{cases} \|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}^2/d^2 \lesssim rd^{-q}\Big(\max\{(\alpha^*)^2,\sigma^2\}\log n\dfrac{\delta d\log d}{n}\Big)^{1-q/2} \\ \|\widehat{\mathbf{\Delta}}\|_{\mathrm{nu}}/d \lesssim rd^{-q}\Big(\max\{\alpha^*,\sigma\}\sqrt{\log n\dfrac{\delta d\log d}{n}}\Big)^{1-q} \end{cases}. \tag{4.18}$$

Remark 3. *Under a specific scaling $\|\mathbf{\Theta}^*\|_{\mathrm{F}} = 1$, $\boldsymbol{X_k} = d\cdot e_ie_j^T$ adopted in [43, 66], our bound for the mean square error $d^{-2}\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}^2$ is equivalent to*

$$\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}^2 \lesssim r\Big(\max\{\alpha(\mathbf{\Theta}^*)^2,\sigma^2\}\frac{d\log d\log n}{n}\Big)^{1-q/2},$$

where $\alpha(\mathbf{\Theta}^*) = \frac{d\|\mathbf{\Theta}^*\|_{\max}}{\|\mathbf{\Theta}^*\|_{\mathrm{F}}} \in [1,d]$ is the spikiness of the desired $\mathbf{\Theta}^*$. Compared with the full-data-based estimator in [66] that achieves near minimax rate, our one-bit estimator only degrades by a minor factor $\log n$, hence is also near minimax. It is quite striking that the underlying matrix can be recovered fairly well based on the one-bit measurements produced by our uniformly dithered quantization.

4.2 Heavy-tailed noise

The heavy-tailed noise is assumed to have bounded second moment in this part, i.e.,

$$\mathbb{E}\epsilon_k = 0,\ \mathbb{E}|\epsilon_k|^2 \le M. \tag{4.19}$$

Although the same notation $\dot{Y}_k$ as sub-Gaussian case is used, note that the one-bit response here is obtained with the truncation step before the dithered quantization. Specifically, Y_k is first truncated to be $\widetilde{Y}_k = \mathrm{sign}(Y_k)\min\{|Y_k|,\eta\}$, then the dithered quantization is applied to obtain $\dot{Y}_k = \mathrm{sign}(\widetilde{Y}_k + \Lambda_k)$ where $\Lambda_k \sim \mathrm{uni}([-\gamma,\gamma])$. The restricted strong convexity is already established by Lemma 4, so it remains to upper bound the right hand side of (4.7) before using Corollary 3.

Lemma 5. *Consider (4.1) under sampling scheme (4.2), max-norm constraint (4.5), and heavy-tailed noise assumption (4.19). For a specific $\delta > 1$, we set the truncation threshold η, dithering scale γ as*

$$\begin{cases} \eta = C_{15}\max\{\alpha^*,\sqrt{M}\}\Big(\dfrac{n}{\delta d\log d}\Big)^{1/4} \\ \gamma = C_{16}\max\{\alpha^*,\sqrt{M}\}\Big(\dfrac{n}{\delta d\log d}\Big)^{1/4} \end{cases}, \tag{4.20}$$

where $C_{16} > C_{15}$, $\gamma > 2\max\{\alpha^*, \sqrt{M}\}$. *If* $\frac{\delta d \log d}{n}$ *is sufficiently small, we have*

$$\Big\|\frac{1}{n}\sum_{k=1}^{n}\big[\,\langle \boldsymbol{X_k}, \boldsymbol{\Theta}^*\rangle - \gamma\cdot \dot{Y}_k\big]\boldsymbol{X_k}\Big\|_{\mathrm{op}} \lesssim \max\{\alpha^*, \sqrt{M}\}\Big(\frac{\delta\log d}{nd^3}\Big)^{1/4} \tag{4.21}$$

with probability higher than $1 - 2d^{1-\delta}$.

Parallel to proof of Theorem 11, a discussion on whether $\widehat{\boldsymbol{\Delta}} \in \mathcal{C}(\psi)$ unfolds some key relations that further lead to the desired error bounds. The result is given in Theorem 12.

Theorem 12. *Under the setting of Lemma 5, assume* $\boldsymbol{\Theta}^*$ *satisfies Assumption 4, we consider the estimator* $\widehat{\boldsymbol{\Theta}}$ *defined in (4.6). Moreover, we set* λ *as*

$$\lambda = C_{17}\max\{\alpha^*, \sqrt{M}\}\Big[\frac{\delta\log d}{nd^3}\Big]^{1/4} \tag{4.22}$$

with sufficiently large C_{17}. *Assume* $\frac{\delta d\log d}{n}$ *is sufficiently small,* $r \gtrsim d^q$, *then with probability at least* $1 - 3d^{1-\delta}$, *we have*

$$\begin{cases} \|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2/d^2 \lesssim rd^{-q}\Big(\max\{(\alpha^*)^2, M\}\sqrt{\dfrac{\delta d\log d}{n}}\Big)^{1-q/2} \\ \|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}}/d \lesssim rd^{-q}\Big(\max\{\alpha^*, \sqrt{M}\}\Big[\dfrac{\delta d\log d}{n}\Big]^{1/4}\Big)^{1-q} \end{cases}. \tag{4.23}$$

Again we can change the bound for mean square error $\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2/d^2$ to the scaling $\|\boldsymbol{\Theta}^*\|_{\mathrm{F}} = 1$ and $\boldsymbol{X_k} = d\cdot e_i e_j^T$, then it reads

$$\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}} \leq \sqrt{r}\Big(\max\{\alpha^*, \sqrt{M}\}\Big[\frac{\delta d\log d}{n}\Big]^{1/4}\Big)^{1-q/2}.$$

This is also consistent with previous two estimation problems where the error rates become essentially slower.Such essential degradation is also due to a trade-off between bias and variance.

To close this section, we point out that our method of one-bit matrix completion is new and essentially different from the existing likelihood approach (see, e.g., [23, 32]). Notably, our method can deal with unknown pre-quantization noise ϵ_k that can be sub-Gaussian or heavy-tailed, while such unknown noise precludes the standard likelihood approach. A review of previous works and more detailed comparison can be found in Appendix D.

5 An Overview of the Proofs

We give an overview of the proofs so that readers can quickly grasp our proof strategies. For illustration we will focus on sub-Gaussian regime. While for heavy-tailed case, we will use two examples to show the same technicalities, together with an optimal trade-off between truncation threshold and dithering scale, can yield the presented results. Finally, we technically compare our work with [43], showing that their techniques cannot apply to our one-bit, heavy-tailed setting.

In this section, we will consider the parameters (n, d, s, r, q) and treated other terms as constants.

For sparse covariance matrix estimation, the element-wise error rate of $\breve{\mathbf{\Sigma}}$ in Theorem 1 is a fundamental element. Unlike the full data case where $\mathbb{E}\big(X_{k,i}X_{k,j}\big) = \sigma^*_{ij}$, $\mathbb{E}\breve{\sigma}_{ij} = \sigma^*_{ij}$ may not hold due to the possibility of $|X_{k,i}| > \gamma$. Thus, we first divide the element-wise error into a concentration term R_1 and a bias term R_2

$$|\breve{\sigma}_{ij} - \sigma^*_{ij}| \le |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| + |\mathbb{E}\breve{\sigma}_{ij} - \sigma^*_{ij}| := R_1 + R_2.$$

Since the quantized data is bounded, a fast concentration rate for R_1 is guaranteed by Hoeffding's inequality, while R_2 can be controlled by standard arguments. We strike a balance between R_1, R_2 by setting $\gamma = O\big(\sqrt{\log\big(\frac{n}{\log d}\big)}\big)$, then the concentration term $R_1 = O\big(\sqrt{\frac{\log d(\log n)^2}{n}}\big)$ dominates the error, hence the error bound only degrades by a factor $\log n$ compared with $O\big(\sqrt{\frac{\log d}{n}}\big)$ for the full-data sample covariance matrix.

Recall that our estimator $\widehat{\mathbf{\Sigma}}$ is defined by element-wisely hard thresholding $\breve{\mathbf{\Sigma}}$, and the procedures to show operator norm error rate of $\widehat{\mathbf{\Sigma}}$ are parallel to corresponding results for the full-data-based hard thresholding estimator $\mathcal{T}_\zeta\big(\sum_{k=1}^n X_k X_k^T/n\big)$ in [20]. In brief, some discussions unfold the element-wise rate $|\widehat{\sigma}_{ij} - \sigma^*_{ij}| = O\big(\min\{|\sigma^*_{ij}|, \sqrt{\frac{\log d(\log n)^2}{n}}\}\big)$, which is tighter than the bound for $|\breve{\sigma}_{ij} - \sigma^*_{ij}|$. This tighter rate, together with the sparsity, can yield a dimension-free bound for the dominating term of operator norm error. Despite a similar proof strategy, we need more involved analyses to deal with some new challenges from the data quantization. These additional efforts, for example, can be seen in the treatment of R_2 (A.8).

For sparse linear regression (including 1-bit QC-CS, 1-bit CS) and matrix completion, we derive the error rates for each problem based on Lemma 2, a framework of trace regression. Compared with the key lemma (Theorem 1) in [43] , we present Lemma 2 in a more general form that accommodates generalized quadratic loss (3.4), and the purpose is that more flexible $\boldsymbol{Q}$, B constructed from the binary data can be used. The advantage of using such framework is a rather clear proof roadmap constituted by two steps:

- *Step 1.* Bound $\|\text{mat}(\boldsymbol{Q}\cdot\text{vec}(\boldsymbol{\Theta}^*)) - \boldsymbol{B}\|_{\text{op}}$ from above and choose λ that guarantee (3.5);
- *Step 2.* Establish the restricted strong convexity (3.7), and invoke (3.8) to obtain the error rate.

We first discuss *Step 1.* In sparse linear regression we need $\lambda \ge 2\|\boldsymbol{Q}\Theta^* - B\|_{\max}$ with some $\boldsymbol{Q}, B$ approximating $\mathbf{\Sigma}_{XX} = \mathbb{E}X_k X_k^T$, $\Sigma_{YX} = \mathbb{E}Y_k X_k$, respectively. Thus, by noting $\mathbf{\Sigma}_{XX}\Theta^* = \Sigma_{YX}$ it can be divided as two approximation error terms

$$\|\boldsymbol{Q}\Theta^* - B\|_{\max} \le \underbrace{\|(\boldsymbol{Q} - \mathbf{\Sigma}_{XX})\Theta^*\|_{\max}}_{\text{approximation term I}} + \underbrace{\|B - \Sigma_{YX}\|_{\max}}_{\text{approximation term II}}.$$

One possibility to control the approximation error term is via existing results. For instance, in 1-bit QC-CS we set $\boldsymbol{Q}$ to be the proposed sparse covariance matrix estimator $\widehat{\mathbf{\Sigma}}$. Thus, the bound of term I follows from results in Section 2 (see, e.g., (B.11)). On the other hand, we can also adopt a standard strategy of bounding the concentration error and the deviation (i.e., bias). For example, we can divide term II into (see, e.g., R_2, R_3 in (B.15))

$$\|B-\Sigma_{YX}\|_{\max} \leq \underbrace{\|B-\mathbb{E}B\|_{\max}}_{\text{concentration term II.1}} + \underbrace{\|\mathbb{E}\big(B-Y_kX_k\big)\|_{\max}}_{\text{bias term II.2}}.$$

For matrix completion the methodology is similar, see (C.1) for example. We apply various concentration inequalities to bound the concentration terms, to name a few, the basic concentration of a sub-Gaussian variable (B.17), Bernstein's inequality (B.16), matrix Bernstein's inequality (C.2). In contrast, more standard tools like Cauchy-Schwarz, Markov's inequality can upper bound bias terms. Let λ_{full} denote the optimal choice of λ in the full-data settings (see, e.g., [28, 43, 65, 66]). As it comes out, in the sub-Gaussian regime of our one-bit setting, one can always strike an almost perfect balance among all the terms such that $\lambda = \text{Poly}(\log n)\cdot\lambda_{\text{full}}$ can guarantee $\lambda \geq 2\|\text{mat}(\boldsymbol{Q}\cdot\text{vec}(\boldsymbol{\Theta}^*)) - \boldsymbol{B}\|_{\text{op}}$ (3.5).

Step 2 concerns the restricted strong convexity of $\boldsymbol{Q}$ with regard to $\widehat{\boldsymbol{\Delta}}$. Note that this mainly hinges on the covariate. Thus, in 1-bit CS and matrix completion where full covariate is available, we can directly borrow existing results from the full-data settings with no quantization [28, 43]. For 1-bit QC-CS with quantized X_k, the desired RSC property straightforwardly follows from $\boldsymbol{\Sigma}_{XX}$'s sparsity and the resulting dimension-free bound of $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_{XX}\|_{\text{op}}$. Finally, we apply (3.8) to obtain the error rates. Since $\lambda = \text{Poly}(\log n)\cdot\lambda_{\text{full}}$ suffices for *Step 1*, under the dithered one-bit quantization scheme, the error rate at worst degrades by logarithmic factor.

In heavy-tailed regime we introduce truncation parameter η and require $\gamma > \eta$. The strategies and technical tools for the proofs are almost the same as sub-Gaussian regime, while the difference is that we can no longer strike a perfect balance among all terms.We briefly give two examples to demonstrate the proofs.

Example 1. (*Theorem 4.*) Similar to sub-Gaussian regime, the element-wise error in Theorem 4 is the basic of Theorem 5, 6. The error can be first divided into two terms as

$$|\breve{\sigma}_{ij} - \sigma^*_{ij}| \leq |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| + |\mathbb{E}(X_{k,i}X_{k,j} - \widetilde{X}_{k,i}\widetilde{X}_{k,j})| := R_1 + R_2.$$

For the concentration term Hoeffding's inequality gives the bound $R_1 = O\big(\gamma^2\sqrt{\frac{\log d}{n}}\big)$. By Cauchy-Schwarz inequality and the assumption of bounded 4-th moment, we show $R_2 = O\big(\frac{1}{\eta^2}\big)$ for the bias term. Intuitively, larger γ brings larger data variance and hence larger R_1, and smaller η corresponds to more biases, and hence larger R_2. Recall that we need $\gamma > \eta$ to ensure the effectiveness of dithering. Hence, we strike a balance between η, γ by setting $\eta, \gamma \asymp \big(\frac{n}{\log d}\big)^{1/8}$, which leads to an overall error $O\big(\sqrt[4]{\frac{\log d}{n}}\big)$. $\triangle$

Example 2. (*Theorem 10.*) To our best knowledge, Theorem 10 presents the first computationally efficient method for 1-bit CS with heavy-tailed sensing vectors, and the rate $O\big(\sqrt[3]{\frac{s^2\log d}{n}}\big)$ (for s-sparse Θ^*) is still faster than the convex approach in [38], which is only for sub-Gaussian regime. Let us start from *Step 1* and first decompose $\|\widehat{\boldsymbol{\Sigma}}_{\tilde{X}\tilde{X}}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max}$ into four terms (see notations given in (3.29))

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max} \le \|(\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} - \mathbf{\Sigma}_{XX})\Theta^*\|_{\max} + \|\widehat{\Sigma}_{YX} - \Sigma_{YX}\|_{\max} := \mathrm{I} + \mathrm{II} \\
&\le \underbrace{\big\|(\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} - \mathbb{E}\widetilde{X}_k\widetilde{X}_k^T)\Theta^*\big\|_{\max}}_{\text{concentration term I.1}} + \underbrace{\big\|\mathbb{E}\big(X_kX_k^T - \widetilde{X}_k\widetilde{X}_k^T\big)\Theta^*\big\|_{\max}}_{\text{bias term I.2}} \\
&\qquad + \underbrace{\big\|\widehat{\Sigma}_{YX} - \mathbb{E}(\gamma\cdot\dot{Y}_k\widetilde{X}_k)\big\|_{\max}}_{\text{concentration term II.1}} + \underbrace{\big\|\mathbb{E}\big(\gamma\cdot\dot{Y}_k\widetilde{X}_k - Y_kX_k\big)\big\|_{\max}}_{\text{bias term II.2}}.
\end{aligned}$$

For two concentration terms, Bernstein's inequality gives $\mathrm{I.1} = O\big(\sqrt{\frac{\log d}{n}} + \eta_X^2\frac{\log d}{n}\big)$ and $\mathrm{II.1} = O\big(\gamma\sqrt{\frac{\log d}{n}} + \frac{\gamma\cdot\eta_X\cdot\log d}{n}\big)$ with probability $1 - d^{-\Omega(1)}$. For two bias terms, some probability arguments and bounded 4-th moment can yield $\mathrm{I.2} = O\big(\frac{1}{\eta_X^2}\big)$ and $\mathrm{II.2} = O\big(\frac{1}{\eta_Y^2} + \frac{1}{\eta_X^2}\big)$. Recall the heavy-tailed Y_k would be quantized to one-bit, we require $\gamma > \eta_Y$. To achieve an optimal trade-off among η_X, η_Y, γ, we set $\eta_X \asymp \big(\frac{n}{\log d}\big)^{1/4}$, $\eta_Y, \gamma \asymp \big(\frac{n}{\log d}\big)^{1/6}$, which gives an overall upper bound $\|\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max} = O\big(\sqrt[3]{\frac{\log d}{n}}\big)$. Hence, $\lambda \asymp \sqrt[3]{\frac{\log d}{n}}$ suffices for $\lambda \ge 2\|\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max}$.

For *Step 2*, since the truncated sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}$ also serves as a plug-in estimator for sparse linear regression in [43], we can borrow their Lemma 2(b) and obtain an error bound. It should be pointed out that if we treat $\widetilde{X}_k$ as data bounded by η_X and deal with I.1, II.1 via Hoeffding's inequality, we can only establish an essentially slower error rate. By contrast, Bernstein's inequality enables us to make full use of X_k's bounded 4-th moment and yield tighter bound. △

To close this section, we compare the heavy-tailed, full-data setting in [43] and our heavy-tailed, one-bit quantized regime from a technical perspective. Although the techniques in [43] can yield (near) minimax rates in sparse linear regression and matrix completion, we point out that they are ineffective in our heavy-tailed regime. Indeed, their key technical ingredients to avoid essential degradation heavily hinge on the moment constraint of the truncated data, which is inherited from the original data. More precisely, the truncated response satisfies $\mathbb{E}|\widetilde{Y}_k|^l \le \mathbb{E}|Y_k|^l$ for any $l > 0$. In contrast, (except for X_k in Theorem 10) heavy-tailed data are further quantized to one-bit in our work, and the quantization evidently ruins the moment constraint since $\mathbb{E}|\gamma\cdot\dot{Y}_k|^l = \gamma^l$. As a consequent, the technical elements in [43] become ineffective in our setting. We give sparse linear regression as a concrete example.

Example 3. (Comparing sparse linear regression in [43] and this work.) In the proof of Lemma 1 in [43], Bernstein's inequality is used to deal with the concentration term $\|\frac{1}{n}\sum_{k=1}^n \widetilde{Y}_k\widetilde{X}_k - \mathbb{E}\widetilde{Y}_k\widetilde{X}_k\|_{\max}$. Thanks to the moment constraints of $\widetilde{X}_k$, $\widetilde{Y}_k$, they can show

$$\Big\|\frac{1}{n}\sum_{k=1}^n \widetilde{Y}_k\widetilde{X}_k - \mathbb{E}\widetilde{Y}_k\widetilde{X}_k\Big\|_{\max} = O\Big(\sqrt{\frac{\log d}{n}} + \frac{\eta_X\eta_Y\log d}{n}\Big) \tag{5.1}$$

with high probability. By contrast, in our Theorem 10 for 1-bit CS, the corresponding term is the concentration term II.1 in Example 2. Since $\gamma\cdot\dot{Y}_k$ fails to inherit the moment constraint from Y_k, the same Bernstein's inequality only delivers (see (B.21), (B.22))

$$\Big\|\frac{1}{n}\sum_{k=1}^n \gamma\cdot\dot{Y}_k\widetilde{X}_k - \mathbb{E}\gamma\cdot\dot{Y}_k\widetilde{X}_k\Big\|_{\max} = O\Big(\gamma\Big(\sqrt{\frac{\log d}{n}} + \frac{\eta_X\log d}{n}\Big)\Big) \tag{5.2}$$

with probability $1-d^{-\Omega(1)}$, which is worse since γ becomes a common factor. Furthermore, in our Theorem 8 for 1-bit QC-CS the corresponding concentration term is $\|\frac{\gamma^2}{n}\sum_{k=1}^n \dot{Y}_k\dot{X}_k - \mathbb{E}\gamma^2\cdot \dot{Y}_k\dot{X}_k\|_{\max}$. Note that both covariate and response are quantized and hence lose the moment constraint, we directly invoke Hoeffding's inequality and obtain (see (B.14), (A.12))

$$\|\frac{\gamma^2}{n}\sum_{k=1}^n \dot{Y}_k\dot{X}_k - \mathbb{E}\gamma^2\cdot \dot{Y}_k\dot{X}_k\|_{\max} = O\big(\gamma^2\sqrt{\frac{\log d}{n}}\big) \tag{5.3}$$

with high probability. This is worse than both (5.1), (5.2), since γ^2 appears as a leading multiplicative factor. It shall be clear that γ or γ^2 appearing as a multiplicative factor of $\sqrt{\frac{\log d}{n}}$ lead to essential degradation. $\triangle$

In matrix completion, similar issue arises in the application of matrix Bernstein's inequality, hence making the techniques developed in [43] useless in our one-bit quantized setting. It is an open question whether our error rates in heavy-tailed regime can be improved.

6 Experimental Results

In this section we present experimental results on synthetic data that can corroborate and demonstrate our theories. To facilitate the presentation flow, the simulation details (e.g., the underlying parameters, covariate) and the algorithms for the convex programming problems are collected in Appendix E.

6.1 Sparse Covariance Matrix Estimation

In our simulation $\mathbf{\Sigma}^*$ has exactly s-sparse columns (see Appendix E).

In sub-Gaussian regime, with high probability Theorem 3 delivers the error bound

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\rm op} \lesssim s\sqrt{\frac{(\log n)^2\log d}{n}}. \tag{6.1}$$

Thus, the operator norm error is expected to only logarithmically depends on the ambient dimension d, while essentially depend on the the sparsity s (that can be viewed as the intrinsic dimension of the problem). We draw X_k from multivariate Gaussian distribution to verify the theory. Specifically, we try $(d,s) = (2500,3), (2700,3), (2900,3), (2700,9)$, and test the sample size $n = 900,\ 1200,\ 1500,\ 1800,\ 2100,\ 2400,\ 2700$ for each (d,s). The (log-log) error curves for all (d,s) are plotted on the left of Figure 1, with the theoretical curve $O\big(\sqrt{\frac{(\log n)^2}{n}}\big)$ also provided for comparison of the error rate. Clearly, the curves with different dimension d but the same sparsity s are almost coincident, which confirms the inessential dependency on d for the error. On the other hand, the estimation error depends on s non-trivially since the curve of $s=9$ is obviously higher, which is consistent with (6.1). Moreover, the experimental curves are roughly parallel to the red one, hence our estimator based on binary data exhibits a near optimal minimax rate.

In heavy-tailed regime, for Σ^* with s-sparse columns Theorem 6 guarantees

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\rm op} \lesssim s\Big(\frac{\log d}{n}\Big)^{1/4}. \tag{6.2}$$

The relation between estimation error and parameters s, d are similar to (6.1), while the convergence rate becomes slower. In our simulations, heavy-tailed data are drawn from Student's t distribution. We test $(d, s) = (2200, 3)$, $(2400, 3)$, $(2600, 3)$, $(2400, 9)$ under sample size

$$n = 900,\ 1200,\ 1500,\ 1800,\ 2100,\ 2400. \tag{6.3}$$

We report the results in the right figure of Figure 1. Consistent with the error bound, three curves with same s but different d are fairly close, while larger s ($s = 9$) leads to essentially larger error. Although our theoretical rate $O\big(n^{-1/4}\big)$ does not match the optimal rate in the classical setting, these curves are well aligned with the theoretical curve, hence the rate is experimentally verified. Furthermore, we test $(d, s) = (2400, 9)$ with the truncation step removed and then show the error curve with legend "no truncation". One shall see the estimation error becomes worse without truncation. Therefore, truncation is not merely of technical importance for proving an error bound, but can indeed improve the estimation in heavy-tailed regime.

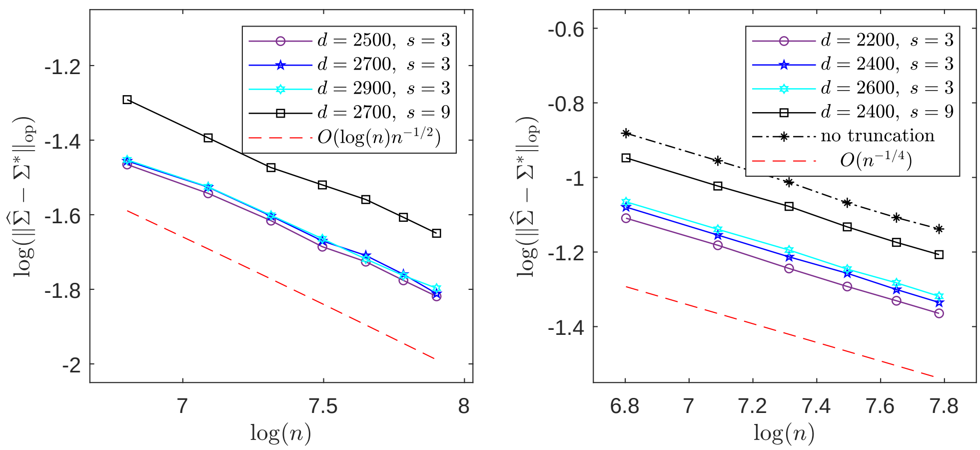

Figure 1: Sparse covariance matrix estimation. Left: Sub-Gaussian; Right: Heavy-tailed.

6.2 Sparse Linear Regression

One-bit quantized-covariate compressed sensing (1-bit QC-CS). In 1-bit QC-CS, both covariate X_k and response Y_k are quantized to one-bit. Note that we use exactly sparse Θ^*, hence in sub-Gaussian regime Theorem 6 delivers the guarantee

$$\|\widehat{\Theta} - \Theta^*\|_2 \lesssim \log n \sqrt{\frac{s \log d}{n}}, \tag{6.4}$$

while for heavy-tailed regime the error bound in Theorem 7 reads as

$$\|\widehat{\Theta} - \Theta^*\|_2 \lesssim \sqrt{s}\Big(\frac{\log d}{n}\Big)^{1/4}. \tag{6.5}$$

With simulation details given in Appendix E, we try Θ^* with $(d, s) = (2400, 3)$, $(2200, 6)$, $(2400, 6)$, $(2600, 6)$ under the same size (6.3). The experimental results in sub-Gaussian regime, heavy-tailed regime are shown as (log-log) curves on the left, the right of Figure 2, respectively. We also plot the theoretical rates for comparison. To show the efficacy of truncation in heavy-tailed regime, keeping other parameters unchanged, we test $(d, s) = (2400, 3)$ without truncation step. The errors are accordingly plot as a curve with legend "no quantization".

The results corroborate the theory from different perspectives. Firstly, the curves with the same s but different d are extremely close, while the errors under $s = 6$ are significantly larger than $s = 3$. This verifies (6.5) and (6.8) that exhibit non-trivial dependence on the sparsity s but only logarithmic dependence on the ambient dimension d. Secondly, since the experimental curves are fairly aligned with the red one, the theoretical convergence rates can be verified. Moreover, comparing the curve of $(d, s) = (2400, 3)$ and "no quantization" on the right of Figure 2, shrinking heavy-tailed data indeed brings better estimation of Θ^*.

One-bit compressed sensing (1-bit CS). Different from the novel setting of 1-bit QC-CS, in 1-bit CS one has full covariate X_k and only quantize Y_k to one-bit. Under s-sparse Θ^*, Theorem 9 gives the near minimax error bound for sub-Gaussian regime

$$\|\widehat{\Theta} - \Theta^*\|_2 \lesssim \sqrt{\frac{s \log d \log n}{n}}, \tag{6.6}$$

while Theorem 12 shows

$$\|\widehat{\Theta} - \Theta^*\|_2 \lesssim \Big(\frac{s^2 \log d}{n}\Big)^{\frac{1}{3}} \tag{6.7}$$

for heavy-tailed regime. To verify the obtained error bounds, under sample size (6.3), in sub-Gaussian regime we try $(d, s) = (2400, 3)$, $(2200, 9)$, $(2400, 9)$, $(2600, 9)$, while in heavy-tailed regime we test $(d, s) = (2400, 3)$, $(2200, 6)$, $(2400, 6)$, $(2600, 6)$. For $(d, s) = (2400, 3)$ with heavy-tailed data, we also conduct an independent simulation with the truncation of X_k, Y_k removed but other conditions unchanged. The (log-log) error curves and the theoretical rates are plotted in Figure 3. Similar to those illustrated in 1-bit QC-CS, several key implications of Figure 3, such as dependence on s, d, can support and demonstrate our theoretical error bounds (6.6), (6.7).

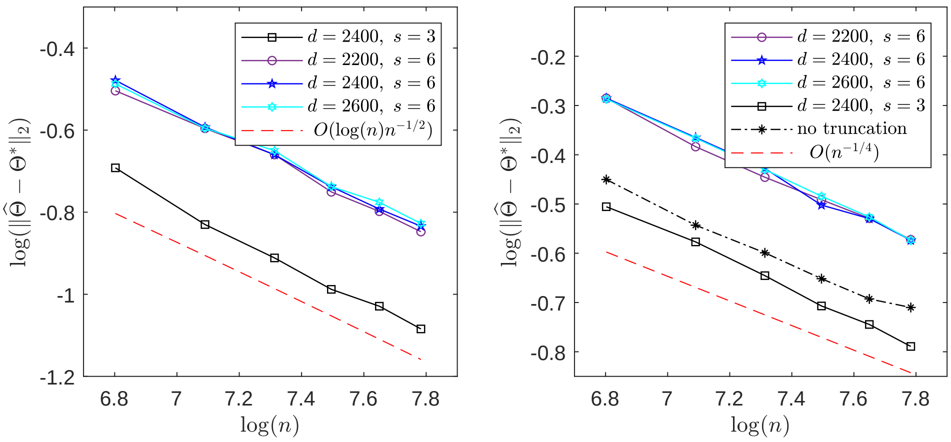

Figure 2: 1-bit QC-CS. Left: Sub-Gaussian; Right: Heavy-tailed.

6.3 Low-rank Matrix Completion

While the error bounds in Theorem 11, 12 are stated with regard to $\frac{\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2}{d^2}$, we first adapt them to our simulation details stated in Appendix E, i.e., the underlying (exactly) low-rank

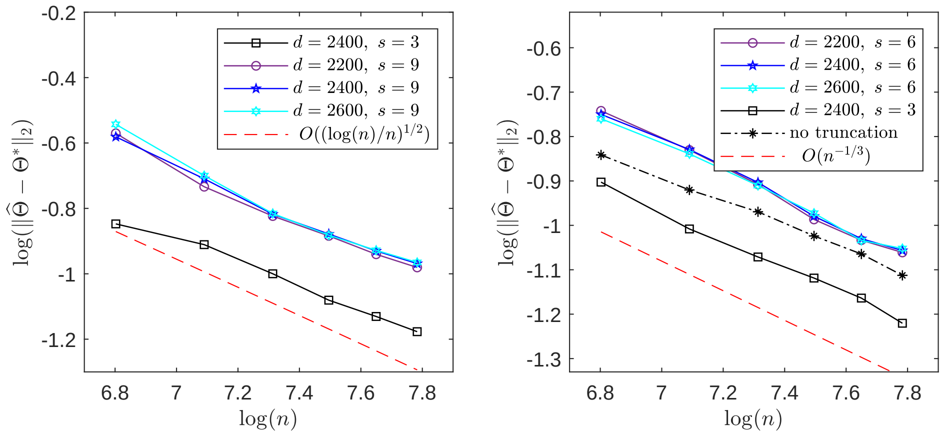

Figure 3: 1-bit CS. Left: Sub-Gaussian; Right: Heavy-tailed.

matrices have comparable spikiness $\alpha(\mathbf{\Theta}^*)$ and unit Frobenius norm, and the noise is moderate compared with the signal. Thus, under sub-Gaussian ϵ_k, we translates MSE error bound in (4.18) into the following via some calculations

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}} \lesssim \sqrt{\frac{rd\log d \log n}{n}}, \tag{6.8}$$

and for heavy-tailed noise (4.23) delivers

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}} \lesssim \Big(\frac{r^2 d\log d}{n}\Big)^{1/4}. \tag{6.9}$$

To corroborate the theoretical error rates, we simulate the proposed one-bit matrix completion method using $\mathbf{\Theta}^*$ with $(d,r) = (100,1)$, $(100,2)$, $(120,2)$, under the sample size $n = 6000$, 7000, 8000, 9000, 10000. In heavy-tailed regime, we also try $(d,r) = (120,2)$ with the response truncation step removed. The experimental results are plotted as (log-log) error curves in Figure 4.

Clearly, in both sub-Gaussian regime (left figure) and heavy-tailed regime (right figure), the errors significantly increase when either r or d becomes larger. This corroborates the implications of (6.8), (6.9) that the estimation error essentially hinges on r and d. Moreover, the experimental curves are well aligned with the theoretical curve, hence the theoretical error rates are observed. Comparing two black curves of $(d,r) = (120,2)$ and "no quantization" in the right figure, the truncation step seems do not bring notable improvement to the recovery of $\mathbf{\Theta}^*$. This is perhaps because the the moderate noise $\frac{1}{250\sqrt{3}} \cdot t(\nu = 3)$ is used in the simulation, thus making the bias-and-variance trade-off less important. On the other hand, we believe a more significant advantage of using the truncation step can be observed under severer noise.

7 Concluding Remarks

In this paper we propose a uniformly dithered one-bit quantization scheme and apply it to three high-dimensional statistical estimation problems, namely sparse covariance matrix estimation, sparse linear regression, and matrix completion. Typical steps for the scheme are truncation, dithering, and quantization. Here, truncation is only applied to heavy-tailed data for robustness, and in dithering a uniformly distributed dithering noise is adopted. From the one-bit

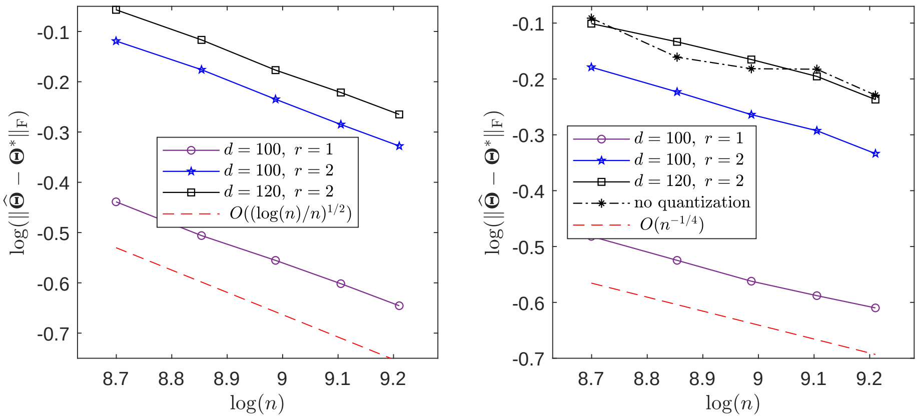

Figure 4: Low-rank matrix completion. Left: Sub-Gaussian; Right: Heavy-tailed.

data, we propose new estimators that are computationally efficient. Under high-dimensional scaling, even only one or two bits per entry are collected, our estimators can recover the underlying parameters fairly well. In sub-Gaussian regime, the proposed estimators achieve near minimax rates, so the quantization costs very little to the recovery. In heavy-tailed regime, the error rates do no match the minimax ones. This degradation is due to a trade-off between bias and variance, see the demonstrations in Subsection 1.3, or the examples in Section 5. However, these results either show essential advantages over comparable papers, or are the first results in such one-bit, heavy-tailed setting. It would be an interesting open question whether the rates in heavy-tailed regime could be faster.

Our work makes considerable contributions to each of the three problems. Compared with [36] that proposed a one-bit covariance matrix estimator $\check{\mathbf{\Sigma}}$, the results presented in Section 2 can be viewed as a two-fold extension, that is, extension to high-dimensional scaling ($n < d$) and to heavy-tailed distribution. We also highlight the technical contribution that bounding higher order moment in (2.11) yields exponentially decaying probability tail in (2.12), see Remark 1. In Section 3, we first propose and study one-bit quantized-covariate compressed sensing (1-bit QC-CS), see Theorem 7, 8. Corresponding results for one-bit compressed sensing (1-bit CS) are also obtained, see Theorem 9, 10. Compared with the 1-bit CS literature, our results exhibit significant improvements from different respects. Most prominently, the sensing vector is not restricted to standard Gaussian, but instead it can be sub-Gaussian or even heavy-tailed with unknown covariance matrix. Besides, some other benefits include faster rate (specifically, even the rate in Theorem 10 is faster than many existing results in 1-bit CS with dithered quantization) and computational feasibility, see the comparison in Appendix D. In Section 4, while all existing papers for one-bit matrix completion (1-bit MC) are in essence based on maximum likelihood estimation, we provide a novel approach that uses a generalized quadratic loss for data fitting. Notably, our new method can handle unknown pre-quantization random noise that precludes the likelihood approach. See a comparison in Appendix D.

Besides the possible improvement of the rates in heavy-tailed regime, there are some other interesting questions for future research. While most assumptions in our work is quite general (e.g., sensing vectors, noise distribution), the dithering noise is restricted to be uniformly distributed. Thus, one may consider the extension to non-uniform dithering noise. This extension seems a bit challenging since the foundational element Lemma 1 heavily relies on a uniformly distributed Λ. Besides, we only consider pre-quantization noise ϵ_k in sparse linear regression and matrix completion. Naturally, one can explore the robustness of our estimators under other types of corruption like post-quantization noise (i.e., sign-flipping). Finally, we point out a more practical aspect. We report that our method may not yield satisfactory estimation in 1-bit MC when entries of $\mathbf{\Theta}^*$ vary a lot in magnitude, which is probably because a common γ can not well preserve the information. It would also be a good question how to make the proposed method more practical for real applications.

References

[1] Albert Ai, Alex Lapanowski, Yaniv Plan, and Roman Vershynin. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

[2] Tuncer C Aysal and Kenneth E Barner. Second-order heavy-tailed distributions and tail analysis. *IEEE transactions on signal processing*, 54(7):2827–2832, 2006.

[3] Youhui Bai, Cheng Li, Quan Zhou, Jun Yi, Ping Gong, Feng Yan, Ruichuan Chen, and Yinlong Xu. Gradient compression supercharged high-performance data parallel dnn training. In *Proceedings of the ACM SIGOPS 28th Symposium on Operating Systems Principles*, pages 359–375, 2021.

[4] Ofer Bar-Shalom and Anthony J Weiss. Doa estimation using one-bit quantized measurements. *IEEE Transactions on Aerospace and Electronic Systems*, 38(3):868–884, 2002.

[5] Richard G Baraniuk, Simon Foucart, Deanna Needell, Yaniv Plan, and Mary Wootters. Exponential decay of reconstruction error from binary measurements of sparse signals. *IEEE Transactions on Information Theory*, 63(6):3368–3385, 2017.

[6] Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

[7] James Bennett, Stan Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, NY, USA., 2007.

[8] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.

[9] Sonia A Bhaskar. Probabilistic low-rank matrix recovery from quantized measurements: Application to image denoising. In *2015 49th Asilomar Conference on Signals, Systems and Computers*, pages 541–545. IEEE, 2015.

[10] Sonia A Bhaskar. Probabilistic low-rank matrix completion from quantized measurements. *The Journal of Machine Learning Research*, 17(1):2131–2164, 2016.

[11] Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[12] Jacob Bien and Robert J Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[13] Atanu Biswas, Sujay Datta, Jason P Fine, and Mark R Segal. *Statistical advances in the biomedical sciences: clinical trials, epidemiology, survival analysis, and bioinformatics*, volume 630. John Wiley & Sons, 2007.

[14] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[15] Petros T Boufounos and Richard G Baraniuk. 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pages 16–21. IEEE, 2008.

[16] Petros T Boufounos, Laurent Jacques, Felix Krahmer, and Rayan Saab. Quantization and compressive sensing. In *Compressed sensing and its applications*, pages 193–237. Springer, 2015.

[17] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[18] T Tony Cai, Zhao Ren, and Harrison H Zhou. Optimal rates of convergence for estimating toeplitz covariance matrices. *Probability Theory and Related Fields*, 156(1-2):101–143, 2013.

[19] T Tony Cai, Cun-Hui Zhang, and Harrison H Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

[20] T Tony Cai and Harrison H Zhou. Minimax estimation of large covariance matrices under ℓ_1-norm. *Statistica Sinica*, pages 1319–1349, 2012.

[21] T Tony Cai and Harrison H Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2012.

[22] Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

[23] Tony Cai and Wen-Xin Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *J. Mach. Learn. Res.*, 14(1):3619–3647, 2013.

[24] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[25] Emmanuel J Candes and Yaniv Plan. A probabilistic and ripless theory of compressed sensing. *IEEE transactions on information theory*, 57(11):7235–7254, 2011.

[26] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[27] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[28] Junren Chen and Michael K Ng. Color image inpainting via robust pure quaternion matrix completion: Error bound and weighted loss. *SIAM Journal on Imaging Sciences*, to appear.

[29] Yongyong Chen, Xiaolin Xiao, and Yicong Zhou. Low-rank quaternion approximation for color image processing. *IEEE Transactions on Image Processing*, 29:1426–1439, 2019.

[30] Yudong Chen and Yuejie Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31, 2018.

[31] Junil Choi, Jianhua Mo, and Robert W Heath. Near maximum-likelihood detector and channel estimator for uplink multiuser massive mimo systems with one-bit adcs. *IEEE Transactions on Communications*, 64(5):2005–2018, 2016.

[32] Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[33] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[34] Oliver De Candido, Hela Jedda, Amine Mezghani, A Lee Swindlehurst, and Josef A Nossek. Reconsidering linear transmit signal processing in 1-bit quantized multi-user miso systems. *IEEE Transactions on Wireless Communications*, 18(1):254–267, 2018.

[35] Sjoerd Dirksen. Quantized compressed sensing: a survey. In *Compressed Sensing and Its Applications*, pages 67–95. Springer, 2019.

[36] Sjoerd Dirksen, Johannes Maly, and Holger Rauhut. Covariance estimation under one-bit quantization. *arXiv preprint arXiv:2104.01280*, 2021.

[37] Sjoerd Dirksen and Shahar Mendelson. Robust one-bit compressed sensing with partial circulant matrices. *arXiv preprint arXiv:1812.06719*, 2018.

[38] Sjoerd Dirksen and Shahar Mendelson. Non-gaussian hyperplane tessellations and robust one-bit compressed sensing. *Journal of the European Mathematical Society*, 23(9):2913–2947, 2021.

[39] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

[40] Noureddine El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, 36(6):2717–2756, 2008.

[41] Jianqing Fan, Wenyan Gong, and Ziwei Zhu. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of econometrics*, 212(1):177–202, 2019.

[42] Jianqing Fan, Kaizheng Wang, Yiqiao Zhong, and Ziwei Zhu. Robust high-dimensional factor models with applications to statistical machine learning. *Statistical Science*, 36(2):303–327, 2021.

[43] Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of statistics*, 49(3):1239, 2021.

[44] Maryam Fazel, Haitham Hindi, and Stephen P Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *Proceedings of the 2003 American Control Conference, 2003.*, volume 3, pages 2156–2162. IEEE, 2003.

[45] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & mathematics with applications*, 2(1):17–40, 1976.

[46] Pengzhi Gao, Ren Wang, Meng Wang, and Joe H Chow. Low-rank matrix recovery from noisy, quantized, and erroneous measurements. *IEEE Transactions on Signal Processing*, 66(11):2918–2932, 2018.

[47] David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.

[48] Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. High dimensional differentially private stochastic optimization with heavy-tailed data. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 227–236, 2022.

[49] Marat Ibragimov, Rustam Ibragimov, and Johan Walden. *Heavy-tailed distributions and robustness in economics and finance*, volume 214. Springer, 2015.

[50] Laurent Jacques, Jason N Laska, Petros T Boufounos, and Richard G Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE transactions on information theory*, 59(4):2082–2102, 2013.

[51] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

[52] Yuan Ke, Stanislav Minsker, Zhao Ren, Qiang Sun, and Wen-Xin Zhou. User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science*, 34(3):454–471, 2019.

[53] Shahin Khobahi, Naveed Naimipour, Mojtaba Soltanalian, and Yonina C Eldar. Deep signal recovery with one-bit quantization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2987–2991. IEEE, 2019.

[54] Alon Kipnis, Yonina C Eldar, and Andrea J Goldsmith. Fundamental distortion limits of analog-to-digital compression. *IEEE Transactions on Information Theory*, 64(9):6013–6033, 2018.

[55] Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[56] Olga Klopp, Jean Lafond, Éric Moulines, and Joseph Salmon. Adaptive multinomial matrix completion. *Electronic Journal of Statistics*, 9(2):2950–2975, 2015.

[57] Olga Klopp, Karim Lounici, and Alexandre B Tsybakov. Robust matrix completion. *Probability Theory and Related Fields*, 169(1):523–564, 2017.

[58] Karin Knudson, Rayan Saab, and Rachel Ward. One-bit compressive sensing with norm estimation. *IEEE Transactions on Information Theory*, 62(5):2748–2758, 2016.

[59] Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[60] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[61] Piotr Kruczek, Radosław Zimroz, and Agnieszka Wyłomańska. How to detect the cyclostationarity in heavy-tailed distributed signals. *Signal Processing*, 172:107514, 2020.

[62] Jean Lafond, Olga Klopp, Eric Moulines, and Joseph Salmon. Probabilistic low-rank matrix completion on finite alphabets. *Advances in Neural Information Processing Systems*, 27, 2014.

[63] Ming Li, Wei Zhao, and Biao Chen. Heavy-tailed prediction error: A difficulty in predicting biomedical signals of noise type. *Computational and Mathematical Methods in Medicine*, 2012, 2012.

[64] Jianhua Mo and Robert W Heath. Limited feedback in single and multi-user mimo systems with finite-bit adcs. *IEEE Transactions on Wireless Communications*, 17(5):3284–3297, 2018.

[65] Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

[66] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

[67] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

[68] Renkun Ni and Quanquan Gu. Optimal statistical and computational rates for one bit matrix completion. In *Artificial Intelligence and Statistics*, pages 426–434. PMLR, 2016.

[69] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2012.

[70] Yaniv Plan and Roman Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

[71] Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

[72] Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with

geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

[73] Mohsen Pourahmadi. *High-dimensional covariance estimation: with high-dimensional data*, volume 882. John Wiley & Sons, 2013.

[74] Jim O Ramsey and Bernard W Silverman. Functional data analysis. *Springer Series in Statistics, New York: Springer Verlag*, 2005.

[75] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

[76] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 813:814, 2015.

[77] Kilian Roth, Jawad Munir, Amine Mezghani, and Josef A Nossek. Covariance based signal parameter estimation of coarse quantized signals. In *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pages 19–23. IEEE, 2015.

[78] Philipp Rütimann and Peter Bühlmann. High dimensional sparse covariance estimation via directed acyclic graphs. *Electronic Journal of Statistics*, 3:1133–1160, 2009.

[79] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*. Citeseer, 2014.

[80] Jie Shen, Pranjal Awasthi, and Ping Li. Robust matrix completion from quantized observations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 397–407. PMLR, 2019.

[81] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

[82] Ananthram Swami and Brian M Sadler. On some detection and estimation problems in heavy-tailed noise. *Signal Processing*, 82(12):1829–1846, 2002.

[83] Christos Thrampoulidis and Ankit Singh Rawat. The generalized lasso for sub-gaussian measurements with dithered quantization. *IEEE Transactions on Information Theory*, 66(4):2487–2500, 2020.

[84] Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

[85] Sara A Van de Geer. *Estimation and testing under sparsity*. Springer, 2016.

[86] Shay Vargaftik, Ran Ben-Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben-Itzhak, and Michael Mitzenmacher. Drive: One-bit distributed mean estimation. *Advances in Neural Information Processing Systems*, 34, 2021.

[87] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[88] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[89] Di Wang and Ruey S Tsay. Robust estimation of high-dimensional vector autoregressive models. *arXiv preprint arXiv:2107.11002*, 2021.

[90] Robert F Woolson and William R Clarke. *Statistical methods for the analysis of biomedical data*, volume 371. John Wiley & Sons, 2011.

[91] Eunho Yang, Aurélie C Lozano, and Pradeep K Ravikumar. Closed-form estimators for high-dimensional generalized linear models. *Advances in Neural Information Processing Systems*, 28, 2015.

[92] Ziwei Zhu and Wenjing Zhou. Taming heavy-tailed features by shrinkage. In *International Conference on Artificial Intelligence and Statistics*, pages 3268–3276. PMLR, 2021.

A Proofs: Sparse Covariance Matrix Estimation

Proof of Lemma 1. Since X and Λ are independent, we have

$$\begin{aligned}&\mathbb{E}\Big[\gamma\cdot\mathrm{sign}(X+\Lambda)\Big]=\mathbb{E}_X\mathbb{E}_\Lambda\Big[\gamma\cdot\mathrm{sign}(X+\Lambda)\Big]=\mathbb{E}_X\Big[\gamma\cdot\mathbb{P}(\Lambda\geq -X)+(-\gamma)\cdot\mathbb{P}(\Lambda<-X)\Big]\\&=\mathbb{E}_X\Big[\gamma\cdot\Big(\frac{\gamma+X}{2\gamma}-\frac{\gamma-X}{2\gamma}\Big)\Big]=\mathbb{E}X.\end{aligned}$$

note that the third equal sign relies on $\gamma\geq B$. □

Proof of Corollary 1. Since Λ_1 and Λ_2 are i.i.d. uniformly distributed on $[-\gamma,\gamma]$ and independent of X, Y, then by using Lemma 1 we have

$$\begin{aligned}&\mathbb{E}\Big[\gamma^2\cdot\mathrm{sign}(X+\Lambda_1)\mathrm{sign}(Y+\Lambda_2)\Big]=\mathbb{E}_{X,Y}\mathbb{E}_{\Lambda_1}\mathbb{E}_{\Lambda_2}\Big[\gamma^2\cdot\mathrm{sign}(X+\Lambda_1)\mathrm{sign}(Y+\Lambda_2)\Big]\\&=\mathbb{E}_{X,Y}\Big(\mathbb{E}_{\Lambda_1}\Big[\gamma\cdot\mathrm{sign}(X+\Lambda_1)\Big]\mathbb{E}_{\Lambda_2}\Big[\gamma\cdot\mathrm{sign}(Y+\Lambda_2)\Big]\Big)=\mathbb{E}XY,\end{aligned}$$

the result follows. □

A.1 Sub-Gaussian Data

Proof of Theorem 1. For fixed i, j, triangle inequality yields

$$|\breve{\sigma}_{ij}-\sigma^*_{ij}|\leq|\breve{\sigma}_{ij}-\mathbb{E}\breve{\sigma}_{ij}|+|\mathbb{E}\breve{\sigma}_{ij}-\sigma^*_{ij}|:=R_1+R_2. \tag{A.1}$$

It suffices to bound R_1, R_2 from above.
<u>*Bound of R_1.*</u> We introduce the element-wise notation of the quantized data as $\dot{X}_{kj}=[\dot{X}_{kj,1},\dot{X}_{kj,2},...,\dot{X}_{kj,d}]^T$, $\forall\, k\in[n], j\in[2]$, then by (2.1) $\breve{\sigma}_{ij}=\frac{1}{n}\sum_{k=1}^n\frac{\gamma^2}{2}\big[\dot{X}_{k1,i}\dot{X}_{k2,j}+\dot{X}_{k2,i}\dot{X}_{k1,j}\big]$. Since $\Big|\frac{\gamma^2}{2}\big[\dot{X}_{k1,i}\dot{X}_{k2,j}+\dot{X}_{k2,i}\dot{X}_{k1,j}\big]\Big|\leq\gamma^2$, Hoeffding's inequality (Proposition 3) yields

$$\mathbb{P}(|\breve{\sigma}_{ij}-\mathbb{E}\breve{\sigma}_{ij}|\geq t)\leq 2\exp(-nt^2/2\gamma^4),\ \forall\, t>0.$$

We set $t = \gamma^2\sqrt{\frac{2\delta\log d}{n}}$ and obtain

$$\mathbb{P}\Big(R_1 \geq \gamma^2\sqrt{\frac{2\delta\log d}{n}}\Big) \leq 2d^{-\delta}. \tag{A.2}$$

<u>*Bound of R_2.*</u> By Corollary 1 and some algebra, we have

$$\begin{aligned} R_2 &= \big|\mathbb{E}\big(\gamma^2\cdot\dot{X}_{k1,i}\dot{X}_{k2,j} - X_{k,i}X_{k,j}\big)\big| \\ &= \Big|\mathbb{E}[\gamma^2\dot{X}_{k1,i}\dot{X}_{k2,j} - X_{k,i}X_{k,j}][\mathbb{1}(\{|X_{k,i}|\geq\gamma\}\cup\{|X_{k,j}|>\gamma\})]\Big| \\ &\leq \mathbb{E}|X_{k,i}X_{k,j}|\mathbb{1}(|X_{k,i}|>\gamma) + \mathbb{E}|X_{k,i}X_{k,j}|\mathbb{1}(|X_{k,j}|>\gamma) := R_{21}+R_{22}. \end{aligned}$$

Note that R_{21}, R_{22} can be bounded likewise, thus we only show the upper bound of R_{21}. We use Cauchy-Schwarz inequality, and then Proposition 1, it yields

$$\begin{aligned} R_{21} &\leq \sqrt{\mathbb{E}|X_{k,i}X_{k,j}|^2}\cdot\sqrt{\mathbb{P}(|X_{k,i}|>\gamma)} \leq \sqrt{\frac{1}{2}\mathbb{E}(|X_{k,i}|^4+|X_{k,j}|^4)}\cdot\sqrt{\mathbb{P}(|X_{k,i}|>\gamma)} \\ &\lesssim \sqrt{\sigma^4}\cdot\sqrt{\exp\big(-\frac{D_1\gamma^2}{\sigma^2}\big)} \leq \sigma^2\exp\big(-\frac{D_1\gamma^2}{2\sigma^2}\big). \end{aligned}$$

We further plug in (2.6) and assume C_1 is sufficiently large such that $D_1C_1^2 \geq 1$, it delivers $R_{21} \lesssim \sigma^2\sqrt{\frac{2\delta\log d}{n}}$. Therefore, we conclude that

$$R_2 = |\mathbb{E}\breve{\sigma}_{ij} - \sigma^*_{ij}| \lesssim \sigma^2\sqrt{\frac{2\delta\log d}{n}}. \tag{A.3}$$

Combining (A.2) and (A.3) we derive $\mathbb{P}\Big(|\breve{\sigma}_{ij}-\sigma^*_{ij}| \lesssim \gamma^2\sqrt{\frac{\delta\log d}{n}}\Big) \geq 1-2d^{\delta}$. With no loss of generality, we can assume $2\delta\log d > e$, then $\gamma^2 \lesssim \sigma^2\log n$, then (2.7) follows. It is not hard to see that (2.8) follows from (2.7) via a union bound. $\square$

Proof of Theorem 2. Since γ has been specified with some C_1, from Theorem 1 we know there exists an absolute constant D_1 such that

$$\mathbb{P}\Big(|\breve{\sigma}_{ij}-\sigma^*_{ij}| \leq D_1\sigma^2\log n\sqrt{\frac{\delta\log d}{n}}\Big) \geq 1-2d^{-\delta}. \tag{A.4}$$

Assume C_2 is sufficiently large such that $C_2 > D_1$. We first rule out $2d^{-\delta}$ probability and assume $|\breve{\sigma}_{ij}-\sigma^*_{ij}| \leq D_1\sigma^2\log n\sqrt{\frac{\delta\log d}{n}}$. Recall that $\widehat{\sigma}_{ij} = \mathcal{T}_\zeta(\breve{\sigma}_{ij})$, we analyse two cases.
<u>*Case 1.*</u> $|\breve{\sigma}_{ij}| < \zeta$, then by definition we have $\widehat{\sigma}_{ij} = 0$, hence $|\widehat{\sigma}_{ij}-\sigma^*_{ij}| = |\sigma^*_{ij}| \leq |\sigma^*_{ij}|$. Besides, by triangle inequality we have $|\sigma^*_{ij}| \leq |\sigma^*_{ij}-\breve{\sigma}_{ij}| + |\breve{\sigma}_{ij}| \leq (D_1+C_2)\sigma^2\log n\sqrt{\frac{\delta\log d}{n}}$, hence we have

$$|\widehat{\sigma}_{ij}-\sigma^*_{ij}| \leq (D_1+C_2+1)\min\Big\{|\sigma^*_{ij}|, \sigma^2\log n\sqrt{\frac{\delta\log d}{n}}\Big\}.$$

<u>*Case 2.*</u> $|\breve{\sigma}_{ij}| \geq \zeta$, then we have $\widehat{\sigma}_{ij} = \breve{\sigma}_{ij}$, hence $|\widehat{\sigma}_{ij}-\sigma^*_{ij}| = |\breve{\sigma}_{ij}-\sigma^*_{ij}| \leq D_1\sigma^2\log n\sqrt{\frac{\delta\log d}{n}}$. Moreover, since $C_2 > D_1$, we have $|\sigma^*_{ij}| \geq |\breve{\sigma}_{ij}| - |\breve{\sigma}_{ij}-\sigma^*_{ij}| \geq (C_2-D_1)\sigma^2\log n\sqrt{\frac{\delta\log d}{n}}$, which

implies that $\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}} \le \frac{1}{C_2 - D_1}|\sigma^*_{ij}|$, hence we have $|\widehat{\sigma}_{ij} - \sigma^*_{ij}| \le \frac{D_1}{C_2-D_1}|\sigma^*_{ij}|$. By putting pieces together we obtain

$$|\widehat{\sigma}_{ij} - \sigma^*_{ij}| \le \Big(D_1 + \frac{D_1}{C_2 - D_1}\Big) \min\Big\{|\sigma^*_{ij}|, \sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\Big\}.$$

Combining two cases leads to (2.10), hence the proof is concluded. □

Proof of Theorem 3. Since γ and ζ are properly set with some C_1, C_2, by Theorem 2, (2.10) holds with some absolute constant D_1 hidden behind "$\lesssim$". For convenience we define

$$\mathscr{A}_{ij} = \Big\{|\widehat{\sigma}_{ij} - \sigma^*_{ij}| \le D_1 \min\{|\sigma^*_{ij}|, \sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\}\Big\}. \tag{A.5}$$

Let $\mathscr{A}^c_{ij}$ be its complement, then we have $\mathbb{P}(\mathscr{A}^c_{ij}) \le 2d^{-\delta}$. For $d \times d$ symmetric matrix $A = [\alpha_1, ..., \alpha_n]$ with columns α_j, we have $\|A\|_{\text{op}} \le \sup_{j\in[d]} \|\alpha_j\|_1$. Thus, some algebra gives

$$\begin{aligned}\mathbb{E}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|^p_{\text{op}} &\le \mathbb{E}\Big[\sup_{j\in[d]}\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\Big]^p \le \mathbb{E}\sup_{j\in[d]}\Big[\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}_{ij}) + \sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}^c_{ij})\Big]^p \\ &\le 2^p\mathbb{E}\sup_{j\in[d]}\Big[\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}_{ij})\Big]^p + 2^p\mathbb{E}\sup_{j\in[d]}\Big[\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}^c_{ij})\Big]^p := R_1 + R_2,\end{aligned} \tag{A.6}$$

<u>*Bound of R_1.*</u> Let us first bound R_1. By (2.3) and (A.5) we have

$$\begin{aligned}\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}_{ij}) &\le \sum_{i=1}^d D_1 \min\Big\{|\sigma^*_{ij}|, \sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\Big\} \\ &\le \sum_{i=1}^d D_1|\sigma^*_{ij}|^q\Big(\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\Big)^{1-q} \le D_1 s\Big(\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\Big)^{1-q}.\end{aligned} \tag{A.7}$$

This further gives

$$R_1 \le \Big(2D_1 s\Big[\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\Big]^{1-q}\Big)^p.$$

<u>*Bound of R_2.*</u> Recall $\widehat{\sigma}_{ij} = \mathcal{T}_\zeta\breve{\sigma}_{ij}$, let $T_1 = \big[\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}^c_{ij})\big]^p$, we have

$$\begin{aligned}T_1 &\le \Big[\sum_{i=1}^d |\sigma^*_{ij}|\mathbb{1}(\mathscr{A}^c_{ij})\mathbb{1}(|\breve{\sigma}_{ij}| < \zeta) + \sum_{i=1}^d |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}|\mathbb{1}(\mathscr{A}^c_{ij}) + \sum_{i=1}^d |\mathbb{E}\breve{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}^c_{ij})\Big]^p \\ &\le (3d)^{p-1}\Big[\sum_{i=1}^d |\sigma^*_{ij}|^p\mathbb{1}(\mathscr{A}^c_{ij})\mathbb{1}(|\breve{\sigma}_{ij}| < \zeta) + \sum_{i=1}^d |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}|^p\mathbb{1}(\mathscr{A}^c_{ij}) + \sum_{i=1}^d |\mathbb{E}\breve{\sigma}_{ij} - \sigma^*_{ij}|^p\mathbb{1}(\mathscr{A}^c_{ij})\Big].\end{aligned}$$

Combining with the form of R_2 yields

$$\begin{aligned}R_2 &\le 2^p\mathbb{E}\sum_{j=1}^d\Big[\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}^c_{ij})\Big]^p \le 6^p d^{p-1}\Big(\mathbb{E}\sum_{i,j} |\sigma^*_{ij}|^p\mathbb{1}(\mathscr{A}^c_{ij})\mathbb{1}(|\breve{\sigma}_{ij}| < \zeta) \\ &+ \mathbb{E}\sum_{i,j} |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}|^p\mathbb{1}(\mathscr{A}^c_{ij}) + \mathbb{E}\sum_{i,j} |\mathbb{E}\breve{\sigma}_{ij} - \sigma^*_{ij}|^p\mathbb{1}(\mathscr{A}^c_{ij})\Big) := 2^p d^{p-1}\big(R_{21} + R_{22} + R_{23}\big).\end{aligned} \tag{A.8}$$

Let us deal with R_{21}, R_{22}, R_{23} separately.
Bound of R_{21}. Suppose the event $\mathscr{A}_{ij}^c \cap \{|\breve{\sigma}_{ij}| < \zeta\}$ holds, then $\widehat{\sigma}_{ij} = 0$, combining with (A.5) we know $|\widehat{\sigma}_{ij} - \sigma_{ij}^*| = |\sigma_{ij}^*| > D_1 \min\{|\sigma_{ij}^*|, \sigma^2 \log n \sqrt{\frac{\delta \log d}{n}}\}$. Recall (A.3), we assume $|\mathbb{E}\breve{\sigma}_{ij} - \sigma_{ij}^*| \le D_{1,0}\sigma^2\sqrt{\frac{\delta \log d}{n}}$ for some constant $D_{1,0}$. To avoid technical complication, we simply assume D_1, C_2 are sufficiently large and satisfy $D_1 \ge \max\{3C_2, 3\}$, $C_2 \ge \max\{D_{1,0}, 10C_1^2\}$.

Combining with (2.9) in Theorem 2, we have

$$|\sigma_{ij}^*| \ge D_1\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}} \ge 3\zeta > 3|\breve{\sigma}_{ij}| \ge 3|\sigma_{ij}^*| - 3|\sigma_{ij}^* - \breve{\sigma}_{ij}|,$$

which implies

$$|\sigma_{ij}^* - \breve{\sigma}_{ij}| \ge \frac{2}{3}|\sigma_{ij}^*| \ \text{ and } \ \zeta \le \frac{1}{3}|\sigma_{ij}^*|. \tag{A.9}$$

Moreover, we have

$$|\mathbb{E}\breve{\sigma}_{ij} - \sigma_{ij}^*| \le D_{1,0}\sigma^2\sqrt{\frac{\delta \log d}{n}} \le \zeta \le \frac{1}{3}|\sigma_{ij}^*|.$$

Besides $|\sigma_{ij}^*| \ge 3\zeta$, based on $|\sigma_{ij}^* - \breve{\sigma}_{ij}| \ge \frac{2}{3}|\sigma_{ij}^*|$, we use triangle inequality and obtain

$$\frac{2}{3}|\sigma_{ij}^*| \le |\sigma_{ij}^* - \breve{\sigma}_{ij}| \le |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| + |\mathbb{E}\breve{\sigma}_{ij} - \sigma_{ij}^*| \le |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| + \frac{1}{3}|\sigma_{ij}^*|, \tag{A.10}$$

which implies $|\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| \ge \frac{1}{3}|\sigma_{ij}^*|$. Therefore, we draw the conclusion that

$$\mathscr{A}_{ij}^c \cap \{|\breve{\sigma}_{ij}| < \zeta\} \Longrightarrow \{|\sigma_{ij}^*| > 3\zeta\} \cap \{|\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| \ge \frac{1}{3}|\sigma_{ij}^*|\}.$$

Now we can invoke Hoeffding's inequality (Proposition 3) and obtain

$$\begin{aligned}
R_{21} &= \sum_{i,j} |\sigma_{ij}^*|^p \mathbb{E}\big[\mathbb{1}(\mathscr{A}_{ij}^c \cap \{|\breve{\sigma}_{ij}| < \zeta\})\big] \\
&\le \sum_{i,j} |\sigma_{ij}^*|^p \mathbb{1}(|\sigma_{ij}^*| > 3\zeta)\mathbb{P}\Big(|\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| \ge \frac{1}{3}|\sigma_{ij}^*|\Big) \\
&\le 2\sum_{i,j} |\sigma_{ij}^*|^p \mathbb{1}(|\sigma_{ij}^*| > 3\zeta)\exp\Big(-\frac{n|\sigma_{ij}^*|^2}{18\gamma^4}\Big)
\end{aligned}$$

Moreover, some calculus can verify $\sup_{y\ge 0} y^{\frac{p}{2}}\exp(-y/36) \le (D_2)^p(\sqrt{p})^p$. Thus, we proceed as

$$\begin{aligned}
&R_{21} \le 2\sum_{i,j} |\sigma_{ij}^*|^p \mathbb{1}(|\sigma_{ij}^*| > 3\zeta)\exp\Big(-\frac{n|\sigma_{ij}^*|^2}{18\gamma^4}\Big) \\
&= 2\Big(\frac{\gamma^2}{\sqrt{n}}\Big)^p \sum_{i,j}\Big(\Big[\frac{n|\sigma_{ij}^*|^2}{\gamma^4}\Big]^{\frac{p}{2}}\exp\Big[-\frac{n|\sigma_{ij}^*|^2}{36\gamma^4}\Big]\Big)\Big(\mathbb{1}(|\sigma_{ij}^*| > 3\zeta)\exp\Big[-\frac{n|\sigma_{ij}^*|^2}{36\gamma^4}\Big]\Big) \\
&\le 2\Big(\frac{\gamma^2}{\sqrt{n}}\Big)^p\Big(\sup_{y\ge 0} y^{\frac{p}{2}}\exp\Big[-\frac{y}{36}\Big]\Big)\Big(d^2\exp\Big[-\frac{n\zeta^2}{4\gamma^4}\Big]\Big),
\end{aligned}$$

Recall (2.6), (2.9) and that we assume $C_2 \geq 10C_1^2$, we have $\zeta \geq 10C_1^2\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}} \geq 10\gamma^2\sqrt{\frac{\delta \log d}{n}}$, which delivers $d^2 \exp(-\frac{n\zeta^2}{4\gamma^4}) \leq d^{2-25\delta}$. We now put pieces together and obtain

$$R_{21} \leq d^{2-25\delta}\Big(2D_2\gamma^2\sqrt{\frac{p}{n}}\Big)^p \leq d^{2-25\delta}\Big(D_3\sigma^2 \log n\sqrt{\frac{\delta}{n}}\Big)^p.$$

<u>*Bound of R_{22}.*</u> By Cauchy-Schwarz inequality we have

$$R_{22} \leq \sum_{i,j}\sqrt{\mathbb{E}|\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}|^{2p}\mathbb{P}(\mathscr{A}_{ij}^c)} \leq \sum_{i,j} d^{-\frac{\delta}{2}}\sqrt{2\mathbb{E}|\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}|^{2p}}.$$

Recall $\breve{\sigma}_{ij} = \frac{1}{n}\sum_{k=1}^n \frac{\gamma^2}{2}\big[\dot{X}_{k1,i}\dot{X}_{k2,j} + \dot{X}_{k2,i}\dot{X}_{k1,j}\big]$ with each summand lying between $[-\frac{\gamma^2}{n}, \frac{\gamma^2}{n}]$, so by Hoeffding's Lemma (e.g., Lemma 1.8 in [76]), $\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}$ is the sum of n independent random variable, and each variable has sub-Gaussian norm scaling $O\big(\frac{\gamma^2}{n}\big)$. Thus, Proposition 2 gives $\|\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}\|_{\psi_2} = O\big(\frac{\gamma^2}{\sqrt{n}}\big)$. Now we invoke Proposition 1(b) to obtain

$$R_{22} \lesssim d^{2-\frac{\delta}{2}}\Big(D_{4,0}\gamma^2\sqrt{\frac{p}{n}}\Big)^p \leq d^{2-\frac{\delta}{2}}\Big(D_4\sigma^2 \log n\sqrt{\frac{\delta}{n}}\Big)^p.$$

<u>*Bound of R_{23}.*</u> Note that $|\mathbb{E}\breve{\sigma}_{ij} - \sigma_{ij}^*|$ is constant, hence we use (A.3) and obtain

$$R_{23} \leq |\mathbb{E}\breve{\sigma}_{ij} - \sigma_{ij}^*|^p \sum_{i,j} 2d^{-\delta} \leq 2d^{2-\delta}\Big(D_5\sigma^2\sqrt{\frac{\delta \log d}{n}}\Big)^p.$$

Now we are in a position to put everything together. By combining the upper bounds for $R_{2i}, i = 1, 2, 3$, we have $R_{21} + R_{22} + R_{23} \leq d^{2-\frac{\delta}{2}}\big(D_6\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\big)^p$. Substitute it into (A.8), recall $p = \frac{\delta}{4}$ and $\delta \geq 4$, we obtain

$$R_2 \leq d^{1-\frac{\delta}{4}}\big(6D_6\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\big)^p \leq \big(6D_6\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\big)^p.$$

This bound is dominated by the bound of R_1 when $\delta \log d(\log n)^2/n$ is sufficiently small (note that conventionally one assumes $s = \Omega(1)$). Thus, there exists absolute constant D_7 such that

$$\mathbb{E}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\text{op}}^p \leq \Big(D_7 s\Big[\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\Big]^{1-q}\Big)^p,$$

which gives (2.11). We further invoke Markov inequality:

$$\begin{aligned}&\mathbb{P}\Big(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\text{op}} \geq e^4 D_7 s\Big[\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\Big]^{1-q}\Big)\\ =&\mathbb{P}\Big(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{\text{op}}^p \geq \Big(e^4 D_7 s\Big[\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\Big]^{1-q}\Big)^p\Big) \leq \exp(-4p) = \exp(-\delta),\end{aligned}$$

(2.12) follows. Now the proof is concluded. □

A.2 Heavy-tailed Data

Proof of Theorem 4. Since $\gamma > \eta \geq |\widetilde{X}_{k,i}|$, by using Corollary 1 we can "expect out" the independent dithering noises $\Gamma_{k1,i}, \Gamma_{k2,j}$,

$$\begin{aligned}&\mathbb{E}\breve{\sigma}_{ij} = \mathbb{E}\big(\gamma^2 \cdot \operatorname{sign}(\widetilde{X}_{k,i} + \Gamma_{k1,i})\operatorname{sign}(\widetilde{X}_{k,j} + \Gamma_{k2,j})\big)\\ =&\mathbb{E}_{\widetilde{X}_{k,i}\widetilde{X}_{k,j}}\Big(\mathbb{E}_{\Gamma_{k1,i}}\big[\gamma \cdot \operatorname{sign}(\widetilde{X}_{k,i} + \Gamma_{k1,i})\big]\Big)\Big(\mathbb{E}_{\Gamma_{k2,j}}\big[\gamma \cdot \operatorname{sign}(\widetilde{X}_{k,j} + \Gamma_{k2,j})\big]\Big) = \mathbb{E}\widetilde{X}_{k,i}\widetilde{X}_{k,j}.\end{aligned}$$

Thus, by triangle inequality we have

$$|\breve{\sigma}_{ij} - \sigma^*_{ij}| \leq |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| + |\mathbb{E}(X_{k,i}X_{k,j} - \widetilde{X}_{k,i}\widetilde{X}_{k,j})| := R_1 + R_2. \tag{A.11}$$

<u>*Bound of R_1*</u>. From $\breve{\sigma}_{ij} = \sum_{k=1}^n \frac{\gamma^2}{2n}(\dot{X}_{k1,i}\dot{X}_{k2,j} + \dot{X}_{k1,j}\dot{X}_{k2,i})$ we know $\breve{\sigma}_{ij}$ is mean of n independent random variables lying in $[-\gamma^2, \gamma^2]$, then by Hoeffding's inequality (Proposition 3) and plug in the value of γ (2.15), we have

$$\mathbb{P}(R_1 \geq t) \leq 2\exp\Big(-\frac{nt^2}{2\gamma^4}\Big) = 2\exp\Big(-\frac{t^2\sqrt{n\delta\log d}}{2C_4^4 M}\Big), \ \forall t > 0. \tag{A.12}$$

Setting $t = \sqrt{2M}C_4^2\big(\frac{\delta\log d}{n}\big)^{1/4}$ yields $\mathbb{P}\big(R_1 \geq \sqrt{2}C_4^2\sqrt{M}\big[\frac{\delta\log d}{n}\big]^{1/4}\big) \leq 2d^{-\delta}$.
<u>*Bound of R_2*</u>. Since the truncated version $\widetilde{X}_{k,i} \neq X_{k,i}$ only when $|X_{k,i}| > \eta$, so

$$\begin{aligned}R_2 &\leq \mathbb{E}\Big[|X_{k,i}X_{k,j} - \widetilde{X}_{k,i}\widetilde{X}_{k,j}|(\mathbb{1}(\{|X_{k,i}| > \eta\} \cup \{|X_{k,j}| > \eta\}))\Big]\\ &= \mathbb{E}\Big[|X_{k,i}X_{k,j}|\mathbb{1}(|X_{k,i}| > \eta)\Big] + \mathbb{E}\Big[|X_{k,i}X_{k,j}|\mathbb{1}(|X_{k,j}| > \eta)\Big] := R_{21} + R_{22}.\end{aligned}$$

By Cauchy-Schwarz inequality, we bound R_{21} by $R_{21} \leq \sqrt{\mathbb{E}|X_{k,i}X_{k,j}|^2\mathbb{P}(|X_{k,i}| > \eta)}$, moreover, we have $\mathbb{E}|X_{k,i}X_{k,j}|^2 \leq \mathbb{E}(|X_{k,i}|^4 + |X_{k,j}|^4)/2 \leq M$. A direct application of Markov inequality yields that $\mathbb{P}(|X_{k,i}| > \eta) \leq \frac{\mathbb{E}|X_{k,i}|^4}{\eta^4} = \frac{M}{\eta^4}$. Plug in the above two inequalities and the value of η, we have $R_{21} \leq \frac{M}{\eta^2} = \frac{\sqrt{M}}{C_3^2}\big(\frac{\delta\log d}{n}\big)^{1/4}$. Since R_{22} can be bounded likewise, it holds that

$$R_2 = |\mathbb{E}(X_{k,i}X_{k,j} - \widetilde{X}_{k,i}\widetilde{X}_{k,j})| \leq \frac{2}{C_3^2}\sqrt{M}\Big(\frac{\delta\log d}{n}\Big)^{\frac{1}{4}}. \tag{A.13}$$

Now we can put things together and obtain (2.16). Moreover, (2.17) follows from a union bound, hence the proof is concluded. □

Proof of Theorem 5. The proof is parallel to that of Theorem 2. For some specified C_3, C_4, by Theorem 4 there exists an absolute constant D_1 such that

$$\mathbb{P}\Big(|\breve{\sigma}_{ij} - \sigma^*_{ij}| \leq D_1\sqrt{M}\Big[\frac{\delta\log d}{n}\Big]^{\frac{1}{4}}\Big) \geq 1 - 2d^{-\delta}. \tag{A.14}$$

We assume $C_5 > D_1$ and first rule out probability $2d^{-\delta}$ in (A.14), so we can proceed the proof upon the event $|\breve{\sigma}_{ij} - \sigma^*_{ij}| \leq D_1\sqrt{M}\big[\frac{\delta\log d}{n}\big]^{1/4}$. According to the threshold ζ we discuss two cases.

Case 1. $|\breve{\sigma}_{ij}| < \zeta$, then we have $\widehat{\sigma}_{ij} = 0$, thus, $|\widehat{\sigma}_{ij} - \sigma^*_{ij}| = |\sigma^*_{ij}| \le |\sigma^*_{ij}|$. Moreover, triangle inequality gives $|\sigma^*_{ij}| \le |\sigma^*_{ij} - \breve{\sigma}_{ij}| + |\breve{\sigma}_{ij}| \le (D_1 + C_5)\sqrt{M}\left(\frac{\delta \log d}{n}\right)^{\frac{1}{4}}$, so we have

$$|\widehat{\sigma}_{ij} - \sigma^*_{ij}| \le (D_1 + C_5 + 1) \min\Big\{|\sigma^*_{ij}|, \sqrt{M}\Big[\frac{\delta \log d}{n}\Big]^{\frac{1}{4}}\Big\}$$

Case 2. $|\breve{\sigma}_{ij}| > \zeta$, then we have $\widehat{\sigma}_{ij} = \breve{\sigma}_{ij}$, which leads to $|\widehat{\sigma}_{ij} - \sigma^*_{ij}| = |\breve{\sigma}_{ij} - \sigma^*_{ij}| \le D_1\sqrt{M}\big[\frac{\delta \log d}{n}\big]^{1/4}$. Let us show it can also be bounded by $|\sigma^*_{ij}|$. A reverse triangle inequality gives

$$|\sigma^*_{ij}| \ge |\breve{\sigma}_{ij}| - |\breve{\sigma}_{ij} - \sigma^*| > \zeta - |\breve{\sigma}_{ij} - \sigma^*| \ge (C_5 - D_1)\sqrt{M}\big[\frac{\delta \log d}{n}\big]^{1/4},$$

so we obtain $\sqrt{M}\big[\frac{\delta \log d}{n}\big]^{1/4} \le \frac{1}{C_5 - D_1}|\sigma^*_{ij}|$. Now we can draw the conclusion that

$$|\widehat{\sigma}_{ij} - \sigma^*_{ij}| \le (D_1 + \frac{D_1}{C_5 - D_1}) \min\Big\{|\sigma^*_{ij}|, \sqrt{M}\Big[\frac{\delta \log d}{n}\Big]^{\frac{1}{4}}\Big\}.$$

Combining two cases leads to (2.20), so we complete the proof. $\square$

Proof of Theorem 6. Since η, γ, ζ are specified with some C_3, C_4, C_5, by Theorem 5 there exists absolute constant D_1 such that (2.20) holds. We define the event

$$\mathscr{A}_{ij} = \Big\{|\widehat{\sigma}_{ij} - \sigma^*_{ij}| \le D_1 \min\{|\sigma^*_{ij}|, \sqrt{M}\big[\frac{\delta \log d}{n}\big]^{\frac{1}{4}}\}\Big\}, \tag{A.15}$$

then we have $\mathbb{P}(\mathscr{A}^c_{ij}) \le 2d^{-\delta}$ (Here, $\mathscr{A}^c_{ij}$ denotes the complementary event). Now we can divide the operator norm error according to $\mathscr{A}_{ij}$ and $\mathscr{A}^c_{ij}$, it gives

$$\mathbb{E}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|^p_{\text{op}} \le \mathbb{E}\Big[\sup_{j\in[d]} \sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\Big]^p \le \mathbb{E}\sup_{j\in[d]}\Big[\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}_{ij}) + \sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}^c_{ij})\Big]^p$$

$$\le 2^p\mathbb{E}\sup_{j\in[d]}\Big[\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}_{ij})\Big]^p + 2^p\mathbb{E}\sup_{j\in[d]}\Big[\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}^c_{ij})\Big]^p := R_1 + R_2.$$

Bound of R_1. By the sparsity (2.3) and (A.15), for any $j \in [d]$ we have

$$\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}_{ij}) \le \sum_{i=1}^d D_1 \min\Big\{|\sigma^*_{ij}|, \sqrt{M}\Big[\frac{\delta \log d}{n}\Big]^{\frac{1}{4}}\Big\}$$

$$\le \sum_{i=1}^d D_1|\sigma^*_{ij}|^q\Big(\sqrt{M}\Big[\frac{\delta \log d}{n}\Big]^{\frac{1}{4}}\Big)^{1-q} \le D_1 s\Big(\sqrt{M}\Big[\frac{\delta \log d}{n}\Big]^{\frac{1}{4}}\Big)^{1-q}.$$

This leads to

$$R_1 \le \Big(2D_1 s M^{(1-q)/2}\big(\frac{\delta \log d}{n}\big)^{(1-q)/4}\Big)^p.$$

Bound of R_2. Let $T_1 = \big[\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}^c_{ij})\big]^p$. Recall that $\widehat{\sigma}_{ij} = \mathcal{T}_\zeta\breve{\sigma}_{ij}$, under $\mathscr{A}^c_{ij}$ we divide the problem into $\{|\breve{\sigma}_{ij}| < \zeta\}$ and $\{|\breve{\sigma}_{ij}| \ge \zeta\}$, then triangle inequality yields

$$T_1 \le \Big[\sum_{i=1}^d |\sigma^*_{ij}|\mathbb{1}(\mathscr{A}^c_{ij})\mathbb{1}(|\breve{\sigma}_{ij}| < \zeta) + \sum_{i=1}^d |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}|\mathbb{1}(\mathscr{A}^c_{ij}) + \sum_{i=1}^d |\mathbb{E}\breve{\sigma}_{ij} - \sigma^*_{ij}|\mathbb{1}(\mathscr{A}^c_{ij})\Big]^p$$

$$\le (3d)^{p-1}\Big[\sum_{i=1}^d |\sigma^*_{ij}|^p\mathbb{1}(\mathscr{A}^c_{ij})\mathbb{1}(|\breve{\sigma}_{ij}| < \zeta) + \sum_{i=1}^d |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}|^p\mathbb{1}(\mathscr{A}^c_{ij}) + \sum_{i=1}^d |\mathbb{E}\breve{\sigma}_{ij} - \sigma^*_{ij}|^p\mathbb{1}(\mathscr{A}^c_{ij})\Big].$$

Now we put it into the expression of R_2 and obtain

$$R_2 \le 2^p \mathbb{E} \sum_{j=1}^d \Big[\sum_{i=1}^d |\widehat{\sigma}_{ij} - \sigma_{ij}^*| \mathbb{1}(\mathscr{A}_{ij}^c) \Big]^p \le 6^p d^{p-1} \Big(\mathbb{E} \sum_{i,j} |\sigma_{ij}^*|^p \mathbb{1}(\mathscr{A}_{ij}^c) \mathbb{1}(|\breve{\sigma}_{ij}| < \zeta)$$
$$+ \mathbb{E} \sum_{i,j} |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}|^p \mathbb{1}(\mathscr{A}_{ij}^c) + \mathbb{E} \sum_{i,j} |\mathbb{E}\breve{\sigma}_{ij} - \sigma_{ij}^*|^p \mathbb{1}(\mathscr{A}_{ij}^c) \Big) := 6^p d^{p-1} (R_{21} + R_{22} + R_{23}). \tag{A.16}$$

<u>*Bound of R_{21}.*</u> Suppose the event $\mathscr{A}_{ij}^c \cap \{|\breve{\sigma}_{ij}| < \zeta\}$ holds, then $\widehat{\sigma}_{ij} = 0$, thus, (A.15) delivers that $|\widehat{\sigma}_{ij} - \sigma_{ij}^*| = |\sigma_{ij}^*| > D_1 \min\{|\sigma_{ij}^*|, \sqrt{M}\big[\frac{\delta \log d}{n}\big]^{1/4}\}$. With no loss of generality, we assume $D_1 \ge \max\{3C_5, 3\}$, and $C_5 \ge \max\{2/C_3^2, 4C_4^2\}$. Combining with (2.19), Theorem 5, we have

$$|\sigma_{ij}^*| \ge D_1 \sqrt{M} \big[\frac{\delta \log d}{n}\big]^{1/4} \ge 3\zeta > 3|\breve{\sigma}_{ij}| \ge 3|\sigma_{ij}^*| - 3|\sigma_{ij}^* - \breve{\sigma}_{ij}|,$$

which implies $|\sigma_{ij}^* - \breve{\sigma}_{ij}| \ge \frac{2}{3}|\sigma_{ij}^*|$ and $\zeta \le \frac{1}{3}|\sigma_{ij}^*|$. Since $\eta < \gamma$, it always holds that $\mathbb{E}\breve{\sigma}_{ij} - \sigma_{ij}^* = \mathbb{E}(X_{k,i}X_{k,j} - \widetilde{X}_{k,i}\widetilde{X}_{k,j})$. Combining (A.13), (2.19) gives

$$|\mathbb{E}\breve{\sigma}_{ij} - \sigma_{ij}^*| \le \frac{2}{C_3^2} \sqrt{M} \Big[\frac{\delta \log d}{n}\Big]^{1/4} \le C_5 \sqrt{M} \Big[\frac{\delta \log d}{n}\Big]^{1/4} \le \zeta \le \frac{1}{3}|\sigma_{ij}^*|.$$

We upper bound $|\sigma_{ij}^* - \breve{\sigma}_{ij}|$ by triangle inequality and have

$$\frac{2}{3}|\sigma_{ij}^*| \le |\sigma_{ij}^* - \breve{\sigma}_{ij}| \le |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| + |\mathbb{E}\breve{\sigma}_{ij} - \sigma_{ij}^*| \le |\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| + \frac{1}{3}|\sigma_{ij}^*|,$$

which implies $|\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| \ge \frac{1}{3}|\sigma_{ij}^*|$. Therefore,

$$\mathscr{A}_{ij}^c \cap \{|\breve{\sigma}_{ij}| < \zeta\} \Longrightarrow \{|\sigma_{ij}^*| > 3\zeta\} \cap \{|\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| \ge \frac{1}{3}|\sigma_{ij}^*|\},$$

so we can bound R_{21} via

$$R_{21} = \sum_{i,j} |\sigma_{ij}^*|^p \mathbb{E}\Big[\mathbb{1}(\mathscr{A}_{ij}^c \cap \{|\breve{\sigma}_{ij}| < \zeta\})\Big]$$
$$\le \sum_{i,j} |\sigma_{ij}^*|^p \mathbb{1}(|\sigma_{ij}^*| > 3\zeta) \mathbb{P}\Big(|\breve{\sigma}_{ij} - \mathbb{E}\breve{\sigma}_{ij}| \ge \frac{1}{3}|\sigma_{ij}^*|\Big) \le 2 \sum_{i,j} |\sigma_{ij}^*|^p \mathbb{1}(|\sigma_{ij}^*| > 3\zeta) \exp\Big(-\frac{n|\sigma_{ij}^*|^2}{18\gamma^4}\Big)$$
$$= 2\Big(\frac{\gamma^2}{\sqrt{n}}\Big)^p \sum_{i,j} \Big(\Big[\frac{n|\sigma_{ij}^*|^2}{\gamma^4}\Big]^{\frac{p}{2}} \exp\Big[-\frac{n|\sigma_{ij}^*|^2}{36\gamma^4}\Big]\Big) \Big(\mathbb{1}(|\sigma_{ij}^*| > 3\zeta) \exp\Big[-\frac{n|\sigma_{ij}^*|^2}{36\gamma^4}\Big]\Big)$$
$$\le 2\Big(\frac{\gamma^2}{\sqrt{n}}\Big)^p \Big(\sup_{y\ge 0} y^{\frac{p}{2}} \exp\Big[-\frac{y}{36}\Big]\Big) \Big(d^2 \exp\Big[-\frac{n\zeta^2}{4\gamma^4}\Big]\Big) \le 2\Big(\frac{\gamma^2}{\sqrt{n}}\Big)^p \Big(\sup_{y\ge 0} y^{\frac{p}{2}} \exp\Big[-\frac{y}{36}\Big]\Big) d^{2-4\delta},$$

where the second inequality is from Hoeffding's inequality (Proposition 3), while we plug in γ, ζ and use $C_5 \ge 4C_4^2$ in the last line. Some calculus show $\sup_{y\ge 0} y^{p/2}\exp(-y/36) \le D_2^p p^{p/2}$. Then we plug in the above inequality and the value of γ (2.15), for some D_3 we have

$$R_{21} \le d^{2-4\delta} \Big(2D_2\gamma^2\sqrt{\frac{p}{n}}\Big)^p \le d^{2-4\delta} \Big(D_3 \sqrt{M} \Big[\frac{\delta}{n \log d}\Big]^{\frac{1}{4}}\Big)^p$$

Bound of R_{22}. This is the same as the corresponding part in the proof of Theorem 3. In brief, we can show an upper bound of the same form, but with different value of γ^2 (given in (2.15)):

$$R_{22} \le d^{2-\frac{\delta}{2}}\Big(D_{4,0}\gamma^2\sqrt{\frac{p}{n}}\Big)^p \le d^{2-\frac{\delta}{2}}\Big(D_4\sqrt{M}\Big[\frac{\delta}{n\log d}\Big]^{\frac{1}{4}}\Big)^p$$

Bound of R_{23}. Note that $|\mathbb{E}\breve{\sigma}_{ij}-\sigma_{ij}^*| = |\mathbb{E}(\widetilde{X}_{k,i}\widetilde{X}_{k,j}-X_{k,i}X_{k,j})|$ is constant which has been bounded in the proof of Theorem 4. In particular, (A.13) gives $|\mathbb{E}\breve{\sigma}_{ij}-\sigma_{ij}^*| \le \frac{2}{C_3^2}\sqrt{M}\big[\frac{\delta\log d}{n}\big]^{1/4}$. By combining with $\mathbb{P}(\mathscr{A}_{ij}^c) \le 2d^{-\delta}$, we bound R_{23} via

$$R_{23} \le |\mathbb{E}\breve{\sigma}_{ij}-\sigma_{ij}^*|^p\sum_{i,j}2d^{-\delta} \le d^{2-\delta}\Big(D_5\sqrt{M}\Big[\frac{\delta\log d}{n}\Big]^{\frac{1}{4}}\Big)^p,$$

Now we are in a position to put things together. By combining the upper bounds for $R_{2i}, i=1,2,3$, we have $R_{21}+R_{22}+R_{23} \le D_6^p d^{2-\frac{\delta}{2}}M^{p/2}\big[\frac{\delta\log d}{n}\big]^{p/4}$. We further substitute it into (A.16), and recall $p=\frac{\delta}{4}$, $\delta\ge 4$, we obtain

$$R_2 \le d^{1-\frac{\delta}{4}}\Big(6D_6\sqrt{M}\Big[\frac{\delta\log d}{n}\Big]^{\frac{1}{4}}\Big)^p \le \Big(6D_6\sqrt{M}\Big[\frac{\delta\log d}{n}\Big]^{\frac{1}{4}}\Big)^p.$$

When $\delta\log d/n$ is small enough, this upper bound for R_2 is smaller than the obtained bound for R_1. Thus, we know there exists absolute constant D_7 such that

$$\mathbb{E}\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\|_{\text{op}}^p \le \Big(D_7 sM^{(1-q)/2}\Big[\frac{\delta\log d}{n}\Big]^{(1-q)/4}\Big)^p,$$

(2.21) follows. We further use Markov inequality:

$$\begin{aligned}&\mathbb{P}\Big(\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\|_{\text{op}} \ge e^4D_7sM^{(1-q)/2}\Big[\frac{\delta\log d}{n}\Big]^{(1-q)/4}\Big)\\ =&\mathbb{P}\Big(\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\|_{\text{op}}^p \ge \Big[e^4D_7sM^{(1-q)/2}\Big[\frac{\delta\log d}{n}\Big]^{(1-q)/4}\Big]^p\Big) \le \exp(-4p)=\exp(-\delta),\end{aligned}$$

this displays (2.22) and concludes the proof. □

B Proofs: Sparse Linear Regression

Proof of Lemma 2. The proof is obtained by modifying and combining Lemma 1 in [67] and Theorem 1 in [43].

I. From the detinition of $\widehat{\mathbf{\Theta}}$ (3.3), we have

$$\mathcal{L}(\widehat{\mathbf{\Theta}})-\mathcal{L}(\mathbf{\Theta}^*) \le \lambda\|\mathbf{\Theta}^*\|_{\text{nu}}-\lambda\|\widehat{\mathbf{\Theta}}\|_{\text{nu}}. \tag{B.1}$$

By (3.4), some algebra delivers that

$$\begin{aligned}\mathcal{L}(\widehat{\mathbf{\Theta}})-\mathcal{L}(\mathbf{\Theta}^*) =& \frac{1}{2}\text{vec}(\widehat{\mathbf{\Delta}})^T\boldsymbol{Q}\text{vec}(\widehat{\mathbf{\Delta}})-\Big\langle\boldsymbol{B},\widehat{\mathbf{\Delta}}\Big\rangle+\text{vec}(\mathbf{\Theta}^*)^T\boldsymbol{Q}\text{vec}(\widehat{\mathbf{\Delta}})\\ =&\frac{1}{2}\text{vec}(\widehat{\mathbf{\Delta}})^T\boldsymbol{Q}\text{vec}(\widehat{\mathbf{\Delta}})+\Big\langle\text{mat}(\boldsymbol{Q}\cdot\text{vec}(\mathbf{\Theta}^*))-\boldsymbol{B},\widehat{\mathbf{\Delta}}\Big\rangle.\end{aligned} \tag{B.2}$$

Since Q is positive semi-definite, combining with $\langle \boldsymbol{A_1}, \boldsymbol{A_2} \rangle \leq \|\boldsymbol{A_1}\|_{\mathrm{op}} \|\boldsymbol{A_2}\|_{\mathrm{nu}}$, (3.5)

$$\mathcal{L}(\widehat{\boldsymbol{\Theta}}) - \mathcal{L}(\boldsymbol{\Theta}^*) \geq -\|\mathrm{mat}(\boldsymbol{Q} \cdot \mathrm{vec}(\boldsymbol{\Theta}^*)) - \boldsymbol{B}\|_{\mathrm{op}} \|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} \geq -\frac{\lambda}{2} \|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}}. \tag{B.3}$$

II. Consider the SVD $\boldsymbol{\Theta}^* = \boldsymbol{U\Sigma V^T} = \begin{bmatrix} \boldsymbol{U_1} & \boldsymbol{U_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma_{11}} & 0 \\ 0 & \boldsymbol{\Sigma_{22}} \end{bmatrix} \begin{bmatrix} \boldsymbol{V_1^T} \\ \boldsymbol{V_2^T} \end{bmatrix}$, where $\boldsymbol{U_1}, \boldsymbol{V_1} \in \mathbb{R}^{d \times z}$, $\boldsymbol{U_2}$, $\boldsymbol{V_2} \in \mathbb{R}^{d \times (d-z)}$ is a partition of singular vectors, $z \in \{0, 1, ..., d\}$ will be specified later. If $z \geq 1$ we consider two linear subspaces of $\mathbb{R}^{d \times d}$ defined as $\mathcal{M} = \{\boldsymbol{U_1 A_1 V_1^*} : \boldsymbol{A_1} \in \mathbb{R}^{z \times z}\}$ and

$$\overline{\mathcal{M}} = \left\{ \begin{bmatrix} \boldsymbol{U_1} & \boldsymbol{U_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{A_1} & \boldsymbol{A_2} \\ \boldsymbol{A_3} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{V_1^T} \\ \boldsymbol{V_2^T} \end{bmatrix} : \boldsymbol{A_1} \in \mathbb{R}^{z \times z}, \boldsymbol{A_2} \in \mathbb{R}^{z \times (d-z)}, \boldsymbol{A_3} \in \mathbb{R}^{(d-z) \times z} \right\},$$

then let $\mathcal{P}_{\mathcal{M}}$ and $\mathcal{P}_{\overline{\mathcal{M}}}$ denote the projection onto $\mathcal{M}$ and $\overline{\mathcal{M}}$ respectively. Given a matrix $\boldsymbol{\Delta} \in \mathbb{R}^{d \times d}$, assume that $\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{U_1} & \boldsymbol{U_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Delta_{11}} & \boldsymbol{\Delta_{12}} \\ \boldsymbol{\Delta_{21}} & \boldsymbol{\Delta_{22}} \end{bmatrix} \begin{bmatrix} \boldsymbol{V_1^T} \\ \boldsymbol{V_2^T} \end{bmatrix}$, then $\mathcal{P}_{\mathcal{M}}$ and $\mathcal{P}_{\overline{\mathcal{M}}}$ have the explicit form

$$\mathcal{P}_{\mathcal{M}} \boldsymbol{\Delta} = \boldsymbol{U_1 \Delta_{11} V_1^T} \text{ and } \mathcal{P}_{\overline{\mathcal{M}}} \boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{U_1} & \boldsymbol{U_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Delta_{11}} & \boldsymbol{\Delta_{12}} \\ \boldsymbol{\Delta_{21}} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{V_1^T} \\ \boldsymbol{V_2^T} \end{bmatrix}.$$

Besides, let $\mathcal{P}_{\mathcal{M}^\perp} \boldsymbol{\Delta} = \boldsymbol{\Delta} - \mathcal{P}_{\mathcal{M}} \boldsymbol{\Delta}$, $\mathcal{P}_{\overline{\mathcal{M}}^\perp} \boldsymbol{\Delta} = \boldsymbol{\Delta} - \mathcal{P}_{\overline{\mathcal{M}}} \boldsymbol{\Delta}$. Note that the nuclear norm is decomposable [67] with respect to the pair of subspaces $(\mathcal{M}, \overline{\mathcal{M}})$ since for any $\boldsymbol{\Delta_1}, \boldsymbol{\Delta_2} \in \mathbb{R}^{d \times d}$, it holds that

$$\|\mathcal{P}_{\mathcal{M}} \boldsymbol{\Delta_1} + \mathcal{P}_{\overline{\mathcal{M}}^\perp} \boldsymbol{\Delta_2}\|_{\mathrm{nu}} = \|\mathcal{P}_{\mathcal{M}} \boldsymbol{\Delta_1}\|_{\mathrm{nu}} + \|\mathcal{P}_{\overline{\mathcal{M}}^\perp} \boldsymbol{\Delta_2}\|_{\mathrm{nu}}. \tag{B.4}$$

By using $\mathcal{P}_{\mathcal{M}}$ and $\mathcal{P}_{\overline{\mathcal{M}}}$, we have $\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} \leq \|\mathcal{P}_{\overline{\mathcal{M}}} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} + \|\mathcal{P}_{\overline{\mathcal{M}}^\perp} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}}$, plug in (B.3) and combine with (B.1), we obtain

$$\|\widehat{\boldsymbol{\Theta}}\|_{\mathrm{nu}} - \|\boldsymbol{\Theta}^*\|_{\mathrm{nu}} \leq \frac{1}{2} \Big[\|\mathcal{P}_{\overline{\mathcal{M}}} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} + \|\mathcal{P}_{\overline{\mathcal{M}}^\perp} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} \Big]. \tag{B.5}$$

In the special case $z = 0$, we just let $\mathcal{P}_{\mathcal{M}} = \mathcal{P}_{\overline{\mathcal{M}}} = \boldsymbol{0}$, it can be easily verified that (B.4), (B.5) and what follow still hold.

III. In this part we derive (3.6). We calculate that

$$\begin{aligned} \|\widehat{\boldsymbol{\Theta}}\|_{\mathrm{nu}} - \|\boldsymbol{\Theta}^*\|_{\mathrm{nu}} &= \|\mathcal{P}_{\mathcal{M}} \boldsymbol{\Theta}^* + \mathcal{P}_{\mathcal{M}^\perp} \boldsymbol{\Theta}^* + \mathcal{P}_{\overline{\mathcal{M}}} \widehat{\boldsymbol{\Delta}} + \mathcal{P}_{\overline{\mathcal{M}}^\perp} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} - \|\mathcal{P}_{\mathcal{M}} \boldsymbol{\Theta}^* + \mathcal{P}_{\mathcal{M}^\perp} \boldsymbol{\Theta}^*\|_{\mathrm{nu}} \\ &\geq \|\mathcal{P}_{\mathcal{M}} \boldsymbol{\Theta}^*\|_{\mathrm{nu}} + \|\mathcal{P}_{\overline{\mathcal{M}}^\perp} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} - \|\mathcal{P}_{\mathcal{M}^\perp} \boldsymbol{\Theta}^*\|_{\mathrm{nu}} - \|\mathcal{P}_{\overline{\mathcal{M}}} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} - \|\mathcal{P}_{\mathcal{M}} \boldsymbol{\Theta}^*\|_{\mathrm{nu}} - \|\mathcal{P}_{\mathcal{M}^\perp} \boldsymbol{\Theta}^*\|_{\mathrm{nu}} \\ &= \|\mathcal{P}_{\overline{\mathcal{M}}^\perp} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} - 2\|\mathcal{P}_{\mathcal{M}^\perp} \boldsymbol{\Theta}^*\|_{\mathrm{nu}} - \|\mathcal{P}_{\overline{\mathcal{M}}} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}}, \end{aligned} \tag{B.6}$$

note that we use decomposability (B.4) and triangle inequality in the third line. By combining (B.5), (B.6) we obtain $\|\mathcal{P}_{\overline{\mathcal{M}}^\perp} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} \leq 3\|\mathcal{P}_{\overline{\mathcal{M}}} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} + 4\|\mathcal{P}_{\mathcal{M}^\perp} \boldsymbol{\Theta}^*\|_{\mathrm{nu}}$, it holds that

$$\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} \leq \|\mathcal{P}_{\overline{\mathcal{M}}^\perp} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} + \|\mathcal{P}_{\overline{\mathcal{M}}} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} \leq 4(\|\mathcal{P}_{\overline{\mathcal{M}}} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} + \|\mathcal{P}_{\mathcal{M}^\perp} \boldsymbol{\Theta}^*\|_{\mathrm{nu}}). \tag{B.7}$$

Assume the singular values of $\boldsymbol{\Theta}^*$ are $\sigma_1(\boldsymbol{\Theta}^*) \geq ... \geq \sigma_d(\boldsymbol{\Theta}^*)$. Instead of choosing z directly we choose a threshold $\tau > 0$ and then let $z = \max\big\{\{0\} \cup \{w \in [d] : \sigma_w(\boldsymbol{\Theta}^*) \geq \tau\}\big\}$. Since $\mathrm{rank}(\mathcal{P}_{\overline{\mathcal{M}}} \widehat{\boldsymbol{\Delta}}) \leq 2z$, we have $\|\mathcal{P}_{\overline{\mathcal{M}}} \widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} \leq \sqrt{2z} \|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}$. Moreover, by (3.2) we have

$$z\tau^q \leq \sum_{k=1}^{z} \sigma_k(\boldsymbol{\Theta}^*)^q \leq \sum_{k=1}^{d} \sigma_k(\boldsymbol{\Theta}^*)^q \leq r,$$

which implies $z \le r\tau^{-q}$. Therefore, we have $\|\mathcal{P}_{\overline{\mathcal{M}}}\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} \le \sqrt{2r}\tau^{-q/2}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}$. By simple algebra we can bound the last term in (B.7) by

$$\|\mathcal{P}_{\mathcal{M}^\perp}\boldsymbol{\Theta}^*\|_{\mathrm{nu}} = \sum_{k=z+1}^{d} \sigma_k(\boldsymbol{\Theta}^*) = \sum_{k=z+1}^{d} \sigma_k(\boldsymbol{\Theta}^*)^q \sigma_k(\boldsymbol{\Theta}^*)^{1-q} \le r\tau^{1-q}.$$

By putting pieces together, we obtain

$$\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} \le 4\Big(\sqrt{2r}\tau^{-\frac{q}{2}}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}} + r\tau^{1-q}\Big), \ \forall \tau > 0.$$

We only consider $\widehat{\boldsymbol{\Delta}} \neq 0$, then we choose $\tau = \left(\frac{\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}}{\sqrt{r}}\right)^{2/(2-q)}$, then we obtain (3.6).

IV. Assume we have RSC (3.7), we derive the convergence rate. With RSC, from (B.2) we have tighter estimation than (B.3):

$$\mathcal{L}(\widehat{\boldsymbol{\Theta}}) - \mathcal{L}(\boldsymbol{\Theta}^*) \ge \frac{1}{2}\kappa\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2 - \frac{\lambda}{2}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}}.$$

On the other hand we have $\mathcal{L}(\widehat{\boldsymbol{\Theta}}) - \mathcal{L}(\boldsymbol{\Theta}^*) \le \lambda\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}}$ from (B.1). By combining them we obtain $\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}} \ge \frac{\kappa}{3\lambda}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2$. Then plug in (3.6), the bound for Frobenius norm in (3.8) follows. Again plug it into (3.6) we obtain the bound for nuclear norm. $\square$

Proof of Corollary 2. (3.9) can be recast as a trace regression $Y_k = \langle \boldsymbol{X_{k,tr}}, \boldsymbol{\Theta^*_{tr}}\rangle + \epsilon_k$, where $\boldsymbol{X_{k,tr}} = \mathrm{diag}(X_k)$, $\boldsymbol{\Theta^*_{tr}} = \mathrm{diag}(\Theta^*)$. Consider the convex set $\mathcal{S} = \{\boldsymbol{\Theta} \in \mathbb{R}^{d\times d} : \boldsymbol{\Theta} = \mathrm{diag}(\Theta), \|\Theta\|_{\max} \le R\}$, let $\boldsymbol{B_{tr}} = \mathrm{diag}(B)$, and $\boldsymbol{Q_{tr}} \in \mathbb{R}^{d^2\times d^2}$ is the matrix whose submatrix constituted of the rows and columns with numbering in $\{1, d+2, 2d+3, ..., d^2\}$ is $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$, and the rows and columns not in $\{1, d+2, 2d+3, ..., d^2\}$ are all zero. Obviously, $\boldsymbol{Q_{tr}}$ is positive semi-definite. It is not hard to see that $\widehat{\Theta}$ defined by (3.11) is equivalent to finding the diagonal matrix $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{tr}}$ via

$$\widehat{\boldsymbol{\Theta}}_{\boldsymbol{tr}} \in \operatorname*{arg\,min}_{\boldsymbol{\Theta}\in\mathcal{S}} \ \mathcal{L}(\boldsymbol{\Theta}) + \lambda\|\boldsymbol{\Theta}\|_{\mathrm{nu}},$$

where the loss function is given by $\mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{2}\mathrm{vec}(\boldsymbol{\Theta})^T\boldsymbol{Q_{tr}}\mathrm{vec}(\boldsymbol{\Theta}) - \langle \boldsymbol{B_{tr}}, \boldsymbol{\Theta}\rangle$, and then let $\widehat{\Theta}$ be the main diagonal of $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{tr}}$. Then all the results follow by using Lemma 2. $\square$

B.1 Sub-Gaussian Data

Proof of Theorem 7. To use Corollary 2 we only need to establish (3.12), (3.14).

I. We first show that when $(\log n)^2 \log d/n$ is sufficiently small, $\widehat{\boldsymbol{\Sigma}}$ is positive definite with high probability. By Assumption 3 and Theorem 3 we have

$$\mathbb{P}\Big(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma_{XX}}\|_{\mathrm{op}} \le D_1\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\Big) \ge 1 - \exp(-\delta). \tag{B.8}$$

Under sufficiently small $(\log n)^2 \log d/n$ we have $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma_{XX}}\|_{\mathrm{op}} \le \kappa_0$ with probability higher than $1 - \exp(-\delta)$. Use $\lambda_{\min}(\cdot)$ to denote the smallest eigenvalue for a symmetric matrix. Combining with $\lambda_{\min}(\boldsymbol{\Sigma}_{XX}) \ge 2\kappa_0$ in Assumption 3, we obtain

$$\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}) \ge \lambda_{\min}(\boldsymbol{\Sigma}_{XX}) - \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_{XX}\|_{\mathrm{op}} \ge \kappa_0, \tag{B.9}$$

which implies that $\widehat{\mathbf{\Sigma}}$ is positive definite, and (3.14) holds.

II. It remains to bound $\|\widehat{\mathbf{\Sigma}}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max}$ and show (3.12) holds with high probability. Let $\Sigma_{YX} = \mathbb{E}Y_kX_k$ and first note that

$$\Sigma_{YX} = \mathbb{E}(Y_kX_k) = \mathbb{E}(X_kX_k^T\Theta^* + \epsilon_kX_k) = \mathbb{E}(X_kX_k^T)\Theta^* = \mathbf{\Sigma}_{XX}\Theta^*.$$

By repeating the proof of Theorem 1, we have the element-wise error for $\widehat{\Sigma}_{YX}$

$$\mathbb{P}\Big(\|\widehat{\Sigma}_{YX} - \Sigma_{YX}\|_{\max} \leq D_2\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}\Big) \geq 1 - 2d^{1-\delta}. \tag{B.10}$$

We now combine (B.8) and (B.10), it holds with probability higher than $1 - 2d^{1-\delta} - \exp(-\delta)$ that

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max} \leq \|\widehat{\mathbf{\Sigma}}\Theta^* - \mathbf{\Sigma}_{XX}\Theta^*\|_{\max} + \|\Sigma_{YX} - \widehat{\Sigma}_{YX}\|_{\max} \\
\leq &\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{XX}\|_{\mathrm{op}}\|\Theta^*\|_2 + \|\Sigma_{YX} - \widehat{\Sigma}_{YX}\|_{\max} \leq (D_1R + D_2)\sigma^2 \log n\sqrt{\frac{\delta \log d}{n}}.
\end{aligned} \tag{B.11}$$

Thus, we can choose sufficiently large C_6 in (3.19) such that $C_6 \geq 2(D_1R + D_2)$, then (3.12) holds with high probability. Now that (3.12) and (3.14) have been verified, Corollary 2 gives (3.15). We further substitute (3.19) into (3.15) and conclude the proof. □

B.2 Heavy-tailed Data

Proof of Theorem 8. The proof is parallel to Theorem 7. By Assumption 3 and Theorem 6, we have the probability tail for operator norm deviation

$$\mathbb{P}\Big(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{XX}\|_{\mathrm{op}} \leq D_1\sqrt{M}\Big[\frac{\delta \log d}{n}\Big]^{1/4}\Big) \geq 1 - \exp(-\delta), \tag{B.12}$$

when $\log d/n$ is sufficiently small, we can assume $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{XX}\|_{\mathrm{op}} \leq \kappa_0$ with probability higher than $1 - \exp(-\delta)$. This, together with $\lambda_{\min}(\mathbf{\Sigma}_{XX}) \geq 2\kappa_0$ given in Assumption 3, gives $\lambda_{\min}(\widehat{\mathbf{\Sigma}}) \geq \kappa_0$ under the same probability. Thus, with high probability $\widehat{\mathbf{\Sigma}}$ is positive definite and (3.14) holds.

It remains to establish (3.12) and apply Corollary 2. By repeating the proof of Theorem 4, we can show the max-norm error for $\widehat{\Sigma}_{YX}$ to approximate $\Sigma_{YX} = \mathbb{E}Y_kX_k$ as

$$\mathbb{P}\Big(\|\widehat{\Sigma}_{YX} - \Sigma_{YX}\|_{\max} \leq D_2\sqrt{M}\Big[\frac{\delta \log d}{n}\Big]^{1/4}\Big) \geq 1 - 2d^{1-\delta}. \tag{B.13}$$

Now we combine (B.12) and (B.13), with probability higher than $1 - \exp(-\delta) - 2d^{1-\delta}$ it yields

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max} \leq \|\widehat{\mathbf{\Sigma}}\Theta^* - \mathbf{\Sigma}_{XX}\Theta^*\|_{\max} + \|\Sigma_{YX} - \widehat{\Sigma}_{YX}\|_{\max} \\
\leq &\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{XX}\|_{\mathrm{op}}\|\Theta^*\|_2 + \|\Sigma_{YX} - \widehat{\Sigma}_{YX}\|_{\max} \leq (D_1R + D_2)\sqrt{M}\Big[\frac{\delta \log d}{n}\Big]^{1/4}.
\end{aligned} \tag{B.14}$$

Thus, in (3.22) we can choose sufficiently large C_7 such that $C_7 \geq 2(D_2R + D_1)$, then we verify $\lambda \geq 2\|\widehat{\mathbf{\Sigma}}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max}$. Now we can use (3.15) in Corollary 2 and plug in (3.22), the desired error bounds follow. □

B.3 One-bit Compressed Sensing

Proof of Theorem 9. We prove the error bound based on Corollary 2. Evidently, we need to show setting $\lambda = C_8\sqrt{\frac{\delta \log d \log n}{n}}$ with sufficiently large C_8 can guarantee $\lambda \geq 2\|\widehat{\mathbf{\Sigma}}_{XX}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max}$. Note that $\mathbf{\Sigma}_{XX}\Theta^* = \mathbb{E}X_k(X_k^T\Theta^*) = \mathbb{E}(Y_kX_k)$, we first invoke triangle inequality to obtain

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}_{XX}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max} &\leq \|\big(\widehat{\mathbf{\Sigma}}_{XX} - \mathbf{\Sigma}_{XX}\big)\Theta^*\|_{\max} + \|\mathbb{E}\big(Y_kX_k - \gamma\cdot\dot{Y}_kX_k\big)\|_{\max} \\ &+ \|\frac{1}{n}\sum_{k=1}^n \gamma\cdot\dot{Y}_kX_k - \mathbb{E}\big(\gamma\cdot\dot{Y}_kX_k\big)\|_{\max} := R_1 + R_2 + R_3.\end{aligned} \tag{B.15}$$

<u>*Bound of R_1.*</u> We first give a standard upper bound of $\|\widehat{\mathbf{\Sigma}}_{XX} - \mathbf{\Sigma}_{XX}\|_{\max}$. For $(i,j)\in[d]\times[d]$, write the i-th entry of X_k as $X_{k,i}$, then we have the (i,j)-th entry of $\widehat{\mathbf{\Sigma}}_{XX} - \mathbf{\Sigma}_{XX}$ is given by $\frac{1}{n}\sum_{k=1}^n X_{k,i}X_{k,j} - \mathbb{E}X_{k,i}X_{k,j}$. Note that $\|X_{k,i}X_{k,j}\|_{\psi_1} \leq \|X_{k,i}\|_{\psi_2}\|X_{k,j}\|_{\psi_2} \leq \sigma^2$, thus by Bernstein's inequality given in Proposition 4 we have

$$\mathbb{P}\Big(\big|\frac{1}{n}\sum_{k=1}^n X_{k,i}X_{k,j} - \mathbb{E}X_{k,i}X_{k,j}\big| \geq t\Big) \leq 2\exp\Big(-D_1\min\big\{\frac{nt^2}{\sigma^4},\frac{nt}{\sigma^2}\big\}\Big),\ \forall\, t>0. \tag{B.16}$$

A union bound further gives

$$\mathbb{P}\big(\|\widehat{\mathbf{\Sigma}}_{XX} - \mathbf{\Sigma}_{XX}\|_{\max} \geq t\big) \leq 2d^2\exp\Big(-D_1\min\big\{\frac{nt^2}{\sigma^4},\frac{nt}{\sigma^2}\big\}\Big),\ \forall\, t>0.$$

Set $t = \max\big\{\frac{1}{D_1},\frac{1}{\sqrt{D_1}}\big\}\sigma^2\sqrt{\frac{\delta\log d}{n}}$, when $n \geq \delta\log d$ we obtain $\|\widehat{\mathbf{\Sigma}}_{XX} - \mathbf{\Sigma}_{XX}\|_{\max} \lesssim \sigma^2\sqrt{\frac{\delta\log d}{n}}$ holds with probability higher than $1-2d^{2-\delta}$. With the same probability, combining $\|\Theta^*\|_1 \leq R$ gives

$$R_1 \leq \|\widehat{\mathbf{\Sigma}}_{XX} - \mathbf{\Sigma}_{XX}\|_{\max}\|\Theta^*\|_1 \lesssim \sigma^2\sqrt{\frac{\delta\log d}{n}}.$$

<u>*Bound of R_2.*</u> By Lemma 1 when $|Y_k| \leq \gamma$ we have $\mathbb{E}_{\Lambda_k}\big(\gamma\cdot\dot{Y}_k\big) = Y_k$. Use this fact and Cauchy-Schwarz inequality, we can first bound R_2 from above as

$$\begin{aligned}R_2 &= \big\|\mathbb{E}\big(Y_k - \gamma\dot{Y}_k\big)X_k\big\|_{\max} = \big\|\mathbb{E}\big(Y_k - \gamma\dot{Y}_k\big)X_k\mathbb{1}\big(|Y_k|>\gamma\big)\big\|_{\max} \\ &\leq \max_{j\in[d]}\ \mathbb{E}\big(|Y_kX_{k,j}|\mathbb{1}(|Y_k|>\gamma)\big) \leq \max_{j\in[d]}\ \sqrt{\mathbb{E}\big[|Y_k|^2|X_{k,j}|^2\big]}\sqrt{\mathbb{P}(|Y_k|>\gamma)} \\ &\leq \max_{j\in[d]}\sqrt{\frac{1}{2}\big(\mathbb{E}|Y_k|^4 + \mathbb{E}|X_{k,j}|^4\big)\mathbb{P}(|Y_k|>\gamma)} \lesssim \sigma^2\exp\big(-\frac{D_2\gamma^2}{\sigma^2}\big),\end{aligned}$$

where the last inequality follows from (3.18) and Proposition 1.
<u>*Bound of R_3.*</u> For $j\in[d]$, by (1.1) it is evident that $\|\gamma\cdot\dot{Y}_kX_{k,j}\|_{\psi_2} \leq \gamma\|X_{k,j}\|_{\psi_2} \leq \gamma\sigma$, hence Proposition 2 further gives $\|\frac{1}{n}\sum_{k=1}^n\gamma\cdot\dot{Y}_kX_{k,j}\|_{\psi_2} \leq \frac{D_3}{\sqrt{n}}\gamma\sigma$. Thus, Proposition 1(a) yields

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{k=1}^n\gamma\cdot\dot{Y}_kX_{k,j} - \mathbb{E}\big(\gamma\cdot\dot{Y}_kX_{k,j}\big)\Big| \geq t\Big) \leq 2\exp\Big(-\frac{D_4nt^2}{\gamma^2\sigma^2}\Big),\ \forall\, t>0. \tag{B.17}$$

Furthermore, a union bound over $j \in [d]$ gives $\mathbb{P}\big(R_3 \geq t\big) \leq 2d \cdot \exp\big(-\frac{D_4 n t^2}{\gamma^2\sigma^2}\big),\ \forall\ t > 0$. We further set $t = \gamma\sigma\sqrt{\frac{\delta \log d}{D_4 n}}$ and obtain

$$\mathbb{P}\Big(R_3 \leq \gamma\sigma\sqrt{\frac{\delta \log d}{D_4 n}}\Big) \geq 1 - 2d^{1-\delta}.$$

By (2.6) we can assume $\sigma < \gamma \lesssim \sigma\sqrt{\log n}$. Thus, by (B.15) and the upper bounds for R_1, R_2, R_3, with probability higher than $1 - 4d^{2-\delta}$ we have

$$\|\widehat{\mathbf{\Sigma}}_{XX}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max} \lesssim \sigma\gamma\sqrt{\frac{\delta \log d}{n}} \lesssim \sigma^2\sqrt{\frac{\delta \log n \log d}{n}}.$$

Since in this Theorem we assume σ is an absolute constant, we can choose sufficiently large C_8 in (3.25) to guarantee $\lambda \geq 2\|\widehat{\mathbf{\Sigma}}_{XX}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max}$ holds with high probability. By Corollary 2, it already leads to (3.13), a relation that facilitates the following discussions.

Now that (3.12) has been verified, we turn to consider the RSC (3.14). When $\frac{\delta \log d}{n}$ is sufficiently small, combining with $\lambda_{\min}(\mathbf{\Sigma}_{XX}) \geq 2\kappa_0$, Lemma 2(a) in [43] gives

$$\mathbb{P}\Big(\widehat{\Delta}^T\widehat{\mathbf{\Sigma}}_{XX}\widehat{\Delta} \geq \kappa_0\|\widehat{\Delta}\|_2^2 - \frac{D_5\delta \log d}{n}\|\widehat{\Delta}\|_1^2\Big) \geq 1 - 3d^{1-\delta}.$$

This event, together with (3.13), implies

$$\widehat{\Delta}^T\widehat{\mathbf{\Sigma}}_{XX}\widehat{\Delta} \geq \kappa_0\|\widehat{\Delta}\|_2^2 - D_6 \cdot \frac{\delta \log d}{n} \cdot s^{\frac{2}{2-q}}\|\widehat{\Delta}\|_2^{\frac{4-4q}{2-q}}. \tag{B.18}$$

We proceed the proof upon the condition (B.18) and divide it into the following two cases.

<u>*Case 1.*</u> If $D_6 \cdot \frac{\delta \log d}{n} \cdot s^{\frac{2}{2-q}}\|\widehat{\Delta}\|_2^{\frac{4-4q}{2-q}} \leq \frac{\kappa_0}{2}\|\widehat{\Delta}\|_2^2$, (B.18) gives the RSC (3.14) with $\kappa = \frac{\kappa_0}{2}$. Thus, we can invoke (3.15) in Corollary 2 and then plug in the value of λ in (3.25). This displays the desired error bounds.
<u>*Case 2.*</u> Otherwise, it holds that

$$D_6 \cdot \frac{\delta \log d}{n} \cdot s^{\frac{2}{2-q}}\|\widehat{\Delta}\|_2^{\frac{4-4q}{2-q}} \geq \frac{\kappa_0}{2}\|\widehat{\Delta}\|_2^2. \tag{B.19}$$

With no loss of generality, we assume $\widehat{\Delta} \neq 0$. Under the scaling that $\sqrt{s}\big(\sqrt{\frac{\delta \log d}{n}}\big)^{1-q/2}$ is sufficiently small we have $q \in (0,1)$ (Since when $q = 0$, $D_6 \cdot \frac{\delta \log d}{n} \cdot s^{\frac{2}{2-q}} < \frac{\kappa_0}{2}$ together with (B.19) gives $\widehat{\Delta} = 0$). Again use sufficiently small $\sqrt{s}\big(\sqrt{\frac{\delta \log d}{n}}\big)^{1-q/2}$, (B.19) delivers

$$\|\widehat{\Delta}\|_2 \lesssim \Big[\sqrt{s}\Big(\sqrt{\frac{\delta \log d}{n}}\Big)^{1-\frac{q}{2}}\Big]^{\frac{2}{q}} \leq \sqrt{s}\Big(\sqrt{\frac{\delta \log d}{n}}\Big)^{1-\frac{q}{2}}.$$

This, together with (3.13), gives the upper bound for $\|\widehat{\Delta}\|_1$ as

$$\|\widehat{\Delta}\|_1 \lesssim s\Big(\sqrt{\frac{\delta \log d}{n}}\Big)^{1-q}.$$

Thus, we conclude the proof. $\square$

Proof of Theorem 10. We follow similar ideas used in the proof of Theorem 9 and intend to invoke Corollary 2. First let us verify the crucial relation $\lambda \geq 2\|\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max}$. Note that $\mathbf{\Sigma}_{XX}\Theta^* = \mathbb{E}Y_kX_k$, by triangle inequality we can divide it into three terms $R_i,\ 1 \leq i \leq 3$

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max} &\leq \big\|\big(\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} - \mathbf{\Sigma}_{XX}\big)\Theta^*\big\|_{\max} + \big\|\mathbb{E}\big(\gamma\cdot\dot{Y}_k\widetilde{X}_k - Y_kX_k\big)\big\|_{\max} \\ &+ \big\|\frac{1}{n}\sum_{k=1}^{n}\gamma\cdot\dot{Y}_k\widetilde{X}_k - \mathbb{E}\big(\gamma\cdot\dot{Y}_k\widetilde{X}_k\big)\big\|_{\max} := R_1 + R_2 + R_3.\end{aligned} \tag{B.20}$$

Bound of R_1. We first decompose $\|\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} - \mathbf{\Sigma}_{XX}\|_{\max}$ as

$$\|\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} - \mathbf{\Sigma}_{XX}\|_{\max} \leq \big\|\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} - \mathbb{E}\widetilde{X}_k\widetilde{X}_k^T\big\|_{\max} + \big\|\mathbb{E}\big(X_kX_k^T - \widetilde{X}_k\widetilde{X}_k^T\big)\big\|_{\max} := R_{11} + R_{12}.$$

Let us deal with them element-wisely. For R_{11} and any $(i,j) \in [d]\times[d]$, recall that the truncated covariate satisfies $|\widetilde{X}_{k,i}| \leq \eta_X$, combining with (3.21) it gives

$$\begin{cases}\sum_{k=1}^{n}\mathbb{E}\big(\widetilde{X}_{k,i}\widetilde{X}_{k,j}\big)^2 \leq \sum_{k=1}^{n}\mathbb{E}X_{k,i}^2X_{k,j}^2 \leq \sum_{k=1}^{n}\frac{1}{2}\big(\mathbb{E}X_{k,i}^4 + \mathbb{E}X_{k,j}^4\big) \leq nM \\ \sum_{k=1}^{n}\mathbb{E}(\widetilde{X}_{k,i}\widetilde{X}_{k,j})_+^q \leq \sum_{k=1}^{n}\mathbb{E}|\widetilde{X}_{k,i}\widetilde{X}_{k,j}|^q \leq (\eta_X^2)^{q-2}\sum_{k=1}^{n}\mathbb{E}(\widetilde{X}_{k,i}\widetilde{X}_{k,j})^2 \leq nM\cdot(\eta_X^2)^{q-2}, \forall q \geq 3\end{cases}.$$

Thus, by the version of Bernstein's inequality given in Theorem 2.10 in [14], we obtain

$$\mathbb{P}\Big(\big|\frac{1}{n}\sum_{k=1}^{n}\widetilde{X}_{k,i}\widetilde{X}_{k,j} - \mathbb{E}\widetilde{X}_{k,i}\widetilde{X}_{k,j}\big| > \sqrt{\frac{2Mt}{n}} + \frac{\eta_X^2 t}{n}\Big) \leq \exp(-t),\ \forall\ t > 0.$$

Moreover, we can use an union bound and get

$$\mathbb{P}\Big(R_{11} > \sqrt{\frac{2Mt}{n}} + \frac{\eta_X^2 t}{n}\Big) \leq d^2\cdot\exp(-t),\ \forall\ t > 0.$$

Thus, we set $t = \delta\log d$ and plug in $\eta_X \asymp \big(\frac{n}{\log d}\big)^{1/4}$, then with probability at least $1 - 2d^{2-\delta}$ we have $R_{11} \lesssim \sqrt{\frac{\delta\log d}{n}}$. We now turn to R_{12} and have the (i,j)-th entry bounded by

$$\big|\mathbb{E}\big(X_{k,i}X_{k,i} - \widetilde{X}_{k,i}\widetilde{X}_{k,j}\big)\big| \leq \mathbb{E}|X_{k,i}X_{k,j}|\big(\mathbb{1}(|X_{k,i}| > \eta_X) + \mathbb{1}(|X_{k,j}| > \eta_X)\big).$$

The two terms can be bounded likewise, so we only deal with one of them by Cauchy-Schwarz inequality and (3.21):

$$\begin{aligned}\mathbb{E}|X_{k,i}X_{k,j}|\mathbb{1}(|X_{k,i}| > \eta_X) &\leq \sqrt{\mathbb{E}|X_{k,i}X_{k,j}|^2}\sqrt{\mathbb{P}(|X_{k,i}| > \eta_X)} \\ &\leq \sqrt{\frac{1}{2}\big(\mathbb{E}|X_{k,i}|^4 + \mathbb{E}|X_{k,j}|^4\big)}\sqrt{\frac{\mathbb{E}|X_{k,i}|^4}{\eta_X^4}} \leq \frac{M}{\eta_X^2} \lesssim \sqrt{\frac{\delta\log d}{n}}.\end{aligned}$$

Therefore, with high probability we have

$$R_1 \leq \|\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} - \mathbf{\Sigma}_{XX}\|_{\max}\|\Theta^*\|_1 \lesssim (R_{11} + R_{12}) \lesssim \sqrt{\frac{\delta\log d}{n}}.$$

Bound of R_2. We consider the j-th entry. Note that $\gamma > \eta_Y$, Lemma 1 gives

$$|\mathbb{E}(\gamma \cdot \dot{Y}_k \widetilde{X}_{k,j} - Y_k X_{k,j})| = |\mathbb{E}(\widetilde{Y}_k \widetilde{X}_{k,j} - Y_k X_{k,j})| \leq \mathbb{E}\big(|Y_k X_{k,j}| \mathbb{1}(|Y_k| > \eta_Y) + \mathbb{1}(|X_{k,j}| > \eta_X)\big).$$

By Cauchy-Schwarz inequality, (3.21) and the value of η_Y, we obtain

$$\mathbb{E}|Y_k X_{k,j}| \mathbb{1}(|Y_k| > \eta_Y) \leq \sqrt{\mathbb{E}|Y_k X_{k,j}|^2 \cdot \mathbb{P}(|Y_k| > \eta_Y)} \leq \frac{M}{\eta_Y^2} \lesssim \Big(\frac{\delta \log d}{n}\Big)^{\frac{1}{3}}.$$

Similarly, it holds that $\mathbb{E}|Y_k X_{k,j}| \mathbb{1}(|X_{k,j} > \eta_X) \leq \frac{M}{\eta_X^2} \lesssim \sqrt{\frac{\delta \log d}{n}}$. Since this is valid for any $j \in [d]$, we obtain $R_2 \lesssim \big(\frac{\delta \log d}{n}\big)^{1/3}$.

Bound of R_3. We consider the j-th entry first. Recall that $|\widetilde{X}_{k,j}| \leq \eta_X$, and by (3.21) we know $\mathbb{E}\widetilde{X}_{k,j}^2 \leq \mathbb{E}|X_{k,j}|^2 \leq \sqrt{M}$, thus we have

$$\begin{cases} \sum_{k=1}^n \mathbb{E}\big(\gamma \cdot \dot{Y}_k \widetilde{X}_{k,j}\big)^2 = \gamma^2 \sum_{k=1}^n \mathbb{E}\widetilde{X}_{k,j}^2 \leq n\sqrt{M}\gamma^2 \\ \sum_{k=1}^n \mathbb{E}(\gamma \cdot \dot{Y}_k \widetilde{X}_{k,j})_+^q \leq \gamma^q \sum_{k=1}^n \mathbb{E}|\widetilde{X}_{k,j}|^q \leq n\sqrt{M}\gamma^2 (\gamma \cdot \eta_X)^{q-2}, \ \forall\, q \geq 3. \end{cases} \tag{B.21}$$

Now, we can invoke the Bernstein's inequality given in Theorem 2.10 in [14] and obtain

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{k=1}^n \gamma \cdot \dot{Y}_k \widetilde{X}_{k,j} - \mathbb{E}\gamma \cdot \dot{Y}_k \widetilde{X}_{k,j}\Big| > \gamma\sqrt{\frac{2\sqrt{M}t}{n}} + \frac{\gamma \cdot \eta_X t}{n}\Big) \leq \exp(-t), \ \forall\, t > 0.$$

Thus, for some absolute constant hidden behind "$\gtrsim$", a union bound gives

$$\mathbb{P}\Big(R_3 \gtrsim \gamma\sqrt{\frac{t}{n}} + \frac{\gamma \cdot \eta_X \cdot t}{n}\Big) \leq d \cdot \exp(-t), \ \forall\, t > 0. \tag{B.22}$$

We set $t = \delta \log d$ and plug in our choices $\eta_X \asymp \big(\frac{n}{\delta \log d}\big)^{1/4}$ and $\gamma \asymp \big(\frac{n}{\delta \log d}\big)^{1/6}$, it yields that $R_3 \lesssim \big(\frac{\delta \log d}{n}\big)^{1/3}$ holds with probability at least $1 - d^{1-\delta}$.

Now combining the upper bounds for $R_i, 1 \leq i \leq 3$ and (B.20), we can choose $\lambda = C_{12}\big(\frac{\delta \log d}{n}\big)^{1/3}$ with sufficiently large C_{12} to guarantee $\lambda \geq 2\|\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}\Theta^* - \widehat{\Sigma}_{YX}\|_{\max}$. Note that $\mathbf{\Sigma}_{XX}\Theta^* = \mathbb{E}Y_k X_k$. By Corollary 2 under the same probability we have (3.13), i.e., $\|\widehat{\Delta}\|_1 \leq 10 s^{\frac{1}{2-q}} \|\widehat{\Delta}\|_2^{\frac{2-2q}{2-q}}$.

To invoke Corollary 2 we still need to establish the RSC (3.14). Note that our choice of the truncation parameter η_X is the same as [43], so we can use Lemma 2(b) therein[2]. Combining with $\lambda_{\min}(\mathbf{\Sigma}_{XX}) \geq 2\kappa_0$ and (3.13), it gives

$$\mathbb{P}\Big(\widehat{\Delta}^T \widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} \widehat{\Delta} \geq 2\kappa_0 \|\widehat{\Delta}\|_2^2 - D_1 \sqrt{\frac{\delta \log d}{n}} s^{\frac{2}{2-q}} \|\widehat{\Delta}\|_2^{\frac{4-4q}{2-q}}\Big) \geq 1 - d^{2-\sqrt{\delta}}.$$

We assume the above event holds, and divide the discussion into two cases.

[2]We mention that this result is presented with the probability term reversed in both Arxiv and Journal

Cases 1. If $D_1\sqrt{\frac{\delta\log d}{n}}s^{\frac{2}{2-q}}\|\widehat{\Delta}\|_2^{\frac{4-4q}{2-q}} \le \kappa_0\|\widehat{\Delta}\|_2^2$, we have $\widehat{\Delta}^T\widehat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}\widehat{\Delta} \ge \kappa_0\|\widehat{\Delta}\|_2^2$, thus confirming the RSC (3.14). Therefore, we can use (3.15) in Corollary 2 and plug in $\lambda \asymp \left(\frac{\delta\log d}{n}\right)^{1/3}$ to yield the error bound $\|\widehat{\Delta}\|_2 \lesssim \sqrt{s}\Big(\frac{\delta\log d}{n}\Big)^{(1-\frac{q}{2})/3} := B_1$.
Cases 2. Otherwise, we assume

$$D_1\sqrt{\frac{\delta\log d}{n}}s^{\frac{2}{2-q}}\|\widehat{\Delta}\|_2^{\frac{4-4q}{2-q}} > \kappa_0\|\widehat{\Delta}\|_2^2. \tag{B.23}$$

With no loss of generality we assume $\widehat{\Delta} \ne 0$. If $q = 0$, under the scaling that $s\big(\sqrt{\frac{\delta\log d}{n}}\big)^{1-q/2}$ is sufficiently small, (B.23) can imply $\widehat{\Delta} = 0$. Thus, we assume $q \in (0,1)$ without losing generality, then (B.23) gives

$$\|\widehat{\Delta}\|_2 \lesssim \Big[s\Big(\sqrt{\frac{\delta\log d}{n}}\Big)^{1-q/2}\Big]^{1/q} = s^{\frac{1}{q}}\Big(\frac{\delta\log d}{n}\Big)^{\frac{1}{2q}-\frac{1}{4}} := B_2.$$

Therefore, we obtain $\|\widehat{\Delta}\|_2 \lesssim \max\{B_1,B_2\} = B_1\max\big\{1,\frac{B_2}{B_1}\big\}$. When $s\big(\sqrt{\frac{\delta\log d}{n}}\big)^{1-q/2}$ is sufficiently small we estimate $\frac{B_2}{B_1}$ as

$$\frac{B_2}{B_1} = s^{\frac{1}{6}}\Big[s\big(\sqrt{\frac{\delta\log d}{n}}\big)^{1-\frac{q}{2}}\Big]^{\frac{1}{q}-\frac{2}{3}} \le s^{\frac{1}{6}}.$$

Therefore, we arrive at the desired upper bound $\|\widehat{\Delta}\|_2 \lesssim s^{\frac{2}{3}}\Big(\frac{\delta\log d}{n}\Big)^{(1-q/2)/3}$. Combining with (3.13), the bound for $\|\widehat{\Delta}\|_1$ follows. $\square$

C Proofs: Low-rank Matrix Completion

C.1 Sub-Gaussian Data

Proof of Lemma 3. **I.** We first prove several facts that would be frequently used later.
Fact 1: $\mathbb{E}\boldsymbol{X_k^TX_k} = \mathbb{E}\boldsymbol{X_kX_k^T} = \boldsymbol{I_d}/d$.
Since X_k and X_k^T follow the same distribution, we only calculate $\mathbb{E}\boldsymbol{X_k^TX_k}$. Equivalent to (4.2) we can assume $\boldsymbol{X_k} = e_{k(i)}e_{k(j)}^T$ where $(k(i),k(j)) \sim \text{uni}([d]\times[d])$. Then we calculate that

$$\begin{aligned}\mathbb{E}\boldsymbol{X_k^TX_k} &= \mathbb{E}_{k(i),k(j)}e_{k(j)}e_{k(i)}^Te_{k(i)}e_{k(j)}^T = \mathbb{E}_{k(j)}e_{k(j)}e_{k(j)}^T\\ &= \sum_{k(j)=1}^d d^{-1}e_{k(j)}e_{k(j)}^T = \boldsymbol{I_d}/d.\end{aligned}$$

Fact 2: Given random matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, then $\|\mathbb{E}\boldsymbol{A}\|_{\text{op}} \le \mathbb{E}\|\boldsymbol{A}\|_{\text{op}}$. Let $\mathcal{S} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$, by using $\|\boldsymbol{B}\|_{\text{op}} = \sup_{U,V\in\mathcal{S}} U^T\boldsymbol{B}V$, we have

$$\|\mathbb{E}\boldsymbol{A}\|_{\text{op}} = \sup_{U,V\in\mathcal{S}}\mathbb{E}[U^T\boldsymbol{A}V] \le \mathbb{E}[\sup_{U,V\in\mathcal{S}}U^T\boldsymbol{A}V] = \mathbb{E}\|\boldsymbol{A}\|_{\text{op}}$$

versions of [43], but the proof therein is fine and can yield what we need here.

Fact 3: Given random matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, then $\|\mathbb{E}(\boldsymbol{A}-\mathbb{E}\boldsymbol{A})^T(\boldsymbol{A}-\mathbb{E}\boldsymbol{A})\|_{\text{op}} \le \|\mathbb{E}\boldsymbol{A^T A}\|_{\text{op}}$, $\|\mathbb{E}(\boldsymbol{A}-\mathbb{E}\boldsymbol{A})(\boldsymbol{A}-\mathbb{E}\boldsymbol{A})^T\|_{\text{op}} \le \|\mathbb{E}\boldsymbol{AA^T}\|_{\text{op}}$.
We only show the first inequality, the second follows likewise. By calculation we have

$$\|\mathbb{E}(\boldsymbol{A}-\mathbb{E}\boldsymbol{A})^T(\boldsymbol{A}-\mathbb{E}\boldsymbol{A})\|_{\text{op}} = \|\mathbb{E}\boldsymbol{A^T A}-\mathbb{E}\boldsymbol{A^T}\mathbb{E}\boldsymbol{A}\|_{\text{op}} \le \|\mathbb{E}\boldsymbol{A^T A}\|_{\text{op}},$$

where we use the positive semi-definiteness of $\mathbb{E}\boldsymbol{A^T}\mathbb{E}\boldsymbol{A}$ and $\mathbb{E}\boldsymbol{A^T A}-\mathbb{E}\boldsymbol{A^T}\mathbb{E}\boldsymbol{A}$.
II. We now start the proof. We first note that

$$\boldsymbol{\Sigma}_{Y\boldsymbol{X}} = \mathbb{E}(Y_k\boldsymbol{X_k}) = \mathbb{E}(\langle \boldsymbol{X_k},\boldsymbol{\Theta}^*\rangle\,\boldsymbol{X_k} + \epsilon_k\boldsymbol{X_k}) = \mathbb{E}(\langle \boldsymbol{X_k},\boldsymbol{\Theta}^*\rangle\,\boldsymbol{X_k}),$$

so by using triangle inequality we obtain

$$\begin{aligned}\Big\|\frac{1}{n}\sum_{k=1}^n\big[\langle \boldsymbol{X_k},\boldsymbol{\Theta}^*\rangle - \gamma\cdot\dot{Y}_k\big]\boldsymbol{X_k}\Big\|_{\text{op}} &\le \Big\|\frac{1}{n}\sum_{k=1}^n\big[\gamma\cdot\dot{Y}_k\boldsymbol{X_k} - \mathbb{E}(\gamma\cdot\dot{Y}_k\boldsymbol{X_k})\big]\Big\|_{\text{op}} + \\ \Big\|\mathbb{E}\big[(\gamma\cdot\dot{Y}_k - Y_k)\boldsymbol{X_k}\big]\Big\|_{\text{op}} &+ \Big\|\frac{1}{n}\sum_{k=1}^n\big[\langle \boldsymbol{X_k},\boldsymbol{\Theta}^*\rangle\,\boldsymbol{X_k} - \mathbb{E}(\langle \boldsymbol{X_k},\boldsymbol{\Theta}^*\rangle\,\boldsymbol{X_k})\big]\Big\|_{\text{op}} := R_1+R_2+R_3.\end{aligned} \tag{C.1}$$

Bound of R_1. We intend to use matrix Bernstein inequality (See Theorem 6.1.1 in [84]) to bound R_1. Consider a finite seqnence of independent, zero-mean random matrices $\big\{\boldsymbol{S_k} := \gamma\cdot\dot{Y}_k\boldsymbol{X_k} - \mathbb{E}(\gamma\cdot\dot{Y}_k\boldsymbol{X_k}) : k\in[n]\big\}$, and by Fact 2 we have

$$\|\boldsymbol{S_k}\|_{\text{op}} \le \|\gamma\cdot\dot{Y}_k\boldsymbol{X_k}\|_{\text{op}} + \|\mathbb{E}(\gamma\cdot\dot{Y}_k\boldsymbol{X_k})\|_{\text{op}} \le \gamma + \mathbb{E}\|\gamma\cdot\dot{Y}_k\boldsymbol{X_k}\|_{\text{op}} \le 2\gamma.$$

Then we bound $\max\{\|n\cdot\mathbb{E}\boldsymbol{S_k S_k^T}\|_{\text{op}}, \|n\cdot\mathbb{E}\boldsymbol{S_k^T S_k}\|_{\text{op}}\}$. By using Fact 1 and Fact 3. we have $\|\mathbb{E}\boldsymbol{S_k S_k^T}\|_{\text{op}} \le \|\mathbb{E}\gamma^2\cdot\boldsymbol{X_k X_k^T}\|_{\text{op}} \le \gamma^2/d$, and similarly it holds that $\|\mathbb{E}\boldsymbol{S_k^T S_k}\|_{\text{op}} \le \gamma^2/d$. Thus, we have

$$\nu\Big(\sum_{k=1}^n\boldsymbol{S_k}\Big) := \max\{\|n\cdot\mathbb{E}\boldsymbol{S_k S_k^T}\|_{\text{op}}, \|n\cdot\mathbb{E}\boldsymbol{S_k^T S_k}\|_{\text{op}}\} \le \frac{n\gamma^2}{d}.$$

By using matrix Bernstein inequality, for any $t>0$ we have

$$\mathbb{P}(R_1\ge t) \le 2d\exp\Big(-\frac{nt^2}{2\gamma[\gamma/d+2t/3]}\Big). \tag{C.2}$$

We let $t = 2\gamma\sqrt{\frac{\delta\log(2d)}{nd}}$, when $\frac{\delta d\log(2d)}{n} < 9/16$ it holds that

$$\mathbb{P}\Big(R_1 \ge 2\gamma\sqrt{\frac{\delta\log(2d)}{nd}}\Big) \le 2d\exp\Big(-\frac{ndt^2}{4\gamma^2}\Big) \le (2d)^{1-\delta}. \tag{C.3}$$

Bound of R_2. We first bound the max norm error $\|\mathbb{E}[(\gamma\cdot\dot{Y}_k - Y_k)\boldsymbol{X_k}]\|_{\max}$. Let $X_{k,ij}$ denotes the (i,j)-th entry of $\boldsymbol{X_k}$. Consider specific $(i,j)\in[d]\times[d]$, then the distribution of $X_{k,ij}$ is given by $\mathbb{P}(X_{k,ij}=1) = d^{-2}$, otherwise $X_{k,ij}=0$. Also, we let Θ^*_{ij} be the (i,j)-th entry of $\boldsymbol{\Theta}^*$. When $|Y_k|<\gamma$ by Lemma 1 we have $\mathbb{E}_{\Lambda_k}\big(\gamma\cdot\dot{Y}_k\big) = Y_k$. Furthermore, it holds that

$$\begin{aligned}&|\mathbb{E}(\gamma\cdot\dot{Y}_k - Y_k)X_{k,ij}| \le \mathbb{E}|Y_k|X_{k,ij}\mathbb{1}(|Y_k|\ge\gamma) = \mathbb{E}\big(\mathbb{E}\big[|Y_k|X_{k,ij}\mathbb{1}(|Y_k|\ge\gamma)\big|X_{k,ij}\big]\big)\\ &= d^{-2}\mathbb{E}\big[|Y_k|X_{k,ij}\mathbb{1}(|Y_k|\ge\gamma)\big|X_{k,ij}=1\big] = d^{-2}\mathbb{E}\big[|\Theta^*_{ij}+\epsilon_k|\mathbb{1}(|\Theta^*_{ij}+\epsilon_k|\ge\gamma)\big].\end{aligned}$$

By (4.5) we have $|\Theta^*_{ij}| \le \alpha^*$, recall that $\gamma \ge 2\alpha^*$, so $|\Theta^*_{ij} + \epsilon_k| \ge \gamma$ implies

$$|\epsilon_k| \ge |\Theta^*_{ij} + \epsilon_k| - |\Theta^*_{ij}| \ge \gamma - \alpha^* \ge \frac{\gamma}{2},$$

which implies $|\epsilon_k| \ge \alpha^*$. Moreover, we obtain

$$|\Theta^*_{ij} + \epsilon_k| \le |\Theta^*_{ij}| + |\epsilon_k| \le \alpha^* + |\epsilon_k| \le 2|\epsilon_k|.$$

Thus, we apply Cauchy-Schwarz inequality, (4.12), Proposition 1, it gives

$$\begin{aligned}
&|\mathbb{E}(\gamma \cdot \dot{Y}_k - Y_k)X_{k,ij}| \le d^{-2}\mathbb{E}\big[|\Theta^*_{ij} + \epsilon_k|\mathbb{1}(|\Theta^*_{ij} + \epsilon_k| \ge \gamma)\big] \le 2d^{-2}\mathbb{E}\big[|\epsilon_k|\mathbb{1}(|\epsilon_k| \ge \frac{\gamma}{2})\big] \\
&\le 2d^{-2}\sqrt{\mathbb{E}\epsilon_k^2}\sqrt{\mathbb{P}\big(|\epsilon_k| \ge \frac{\gamma}{2}\big)} \lesssim d^{-2}\sigma \exp\Big(-\frac{D_1\gamma^2}{\sigma^2}\Big) \lesssim d^{-1}\sigma\sqrt{\frac{\delta\log(2d)}{nd}},
\end{aligned}$$

where the last "$\lesssim$" follows from γ given in (4.12) with sufficiently large C_{13}. Since the estimation holds for any (i,j), we have $\|\mathbb{E}[(\gamma\dot{Y}_k - Y_k)\boldsymbol{X_k}]\|_{\max} \le d^{-1}\sigma\sqrt{\frac{\delta\log(2d)}{nd}}$. By the relation between $\|.\|_{\text{op}}$ and $\|.\|_{\max}$, it further gives

$$R_2 \le d \cdot \|\mathbb{E}[(\gamma \cdot \dot{Y}_k - Y_k)\boldsymbol{X_k}]\|_{\max} \lesssim \sigma\sqrt{\frac{\delta\log(2d)}{nd}}.$$

Bound of R_3. Similar to R_1 we use matrix Bernstein inequality. We consider the finite independent, zero-mean random matrix sequence

$$\big\{\boldsymbol{W_k} := \langle \boldsymbol{X_k}, \boldsymbol{\Theta^*}\rangle \boldsymbol{X_k} - \mathbb{E}(\langle \boldsymbol{X_k}, \boldsymbol{\Theta^*}\rangle \boldsymbol{X_k}) : k \in [n]\big\}.$$

Note that $|\langle \boldsymbol{X_k}, \boldsymbol{\Theta^*}\rangle| \le \alpha^*$, so $\|\langle \boldsymbol{X_k}, \boldsymbol{\Theta^*}\rangle \boldsymbol{X_k}\|_{\text{op}} \le \alpha^*$, by Fact 2 we have

$$\|\boldsymbol{W_k}\|_{\text{op}} \le \|\langle \boldsymbol{X_k}, \boldsymbol{\Theta^*}\rangle \boldsymbol{X_k}\|_{\text{op}} + \|\mathbb{E}\langle \boldsymbol{X_k}, \boldsymbol{\Theta^*}\rangle \boldsymbol{X_k}\|_{\text{op}} \le 2\alpha^* \le 2\gamma.$$

By using Fact 1 and Fact 3, we obtain $\|\mathbb{E}\boldsymbol{W_k}\boldsymbol{W_k^T}\|_{\text{op}} \le \|\mathbb{E}\langle \boldsymbol{X_k}, \boldsymbol{\Theta^*}\rangle^{\boldsymbol{2}} \boldsymbol{X_k}\boldsymbol{X_k^T}\|_{\text{op}} \le \frac{(\alpha^*)^2}{d}$. Likewise we have $\|\mathbb{E}\boldsymbol{W_k^T}\boldsymbol{W_k}\|_{\text{op}} \le \frac{(\alpha^*)^2}{d}$, so we derive the bound

$$\nu\Big(\sum_{k=1}^n \boldsymbol{W_k}\Big) := \max\{\|n \cdot \mathbb{E}\boldsymbol{W_k}\boldsymbol{W_k^T}\|_{\text{op}}, \|n \cdot \mathbb{E}\boldsymbol{W_k^T}\boldsymbol{W_k}\|_{\text{op}}\} \le \frac{n(\alpha^*)^2}{d} \le \frac{n\gamma^2}{d}.$$

Parallel to R_1, by using Matrix Bernstein inequality and set $t = 2\gamma\sqrt{\frac{\delta\log(2d)}{nd}}$,

$$\mathbb{P}\Big(R_3 \ge 2\gamma\sqrt{\frac{\delta\log(2d)}{nd}}\Big) \le (2d)^{1-\delta}. \tag{C.4}$$

We combine the obtained upper bounds for R_1, R_2, R_3 and draw the conclusion that with probability higher than $1 - 2d^{1-\delta}$, we have

$$\Big\|\frac{1}{n}\sum_{k=1}^n \big[\langle \boldsymbol{X_k}, \boldsymbol{\Theta^*}\rangle - \gamma \cdot \dot{Y}_k\big]\boldsymbol{X_k}\Big\|_{\text{op}} \lesssim \gamma\sqrt{\frac{\delta\log d}{nd}}.$$

Now we can use $\gamma \le C_{13}\max\{\alpha^*, \sigma\}\sqrt{\log n}$ to conclude the proof. $\square$

Proof of Lemma 4. **I.** We first decompose the complementary event of (4.15) which can be stated as $\mathscr{B} = \{\exists\ \mathbf{\Theta_0} \in \mathcal{C}(\psi), \text{s.t. } \mathcal{F}_{\mathscr{X}}(\mathbf{\Theta_0}) \le \kappa d^{-2}\|\mathbf{\Theta_0}\|_{\rm F}^2 - T_0\}$. Note that $\mathbb{E}\mathcal{F}_{\mathscr{X}}(\mathbf{\Theta}) = d^{-2}\|\mathbf{\Theta}\|_{\rm F}^2$, so $\mathscr{B}$ implies the following event

$$\{\exists\ \mathbf{\Theta_0} \in \mathcal{C}(\psi), \text{s.t. } |\mathcal{F}_{\mathscr{X}}(\mathbf{\Theta_0}) - \mathbb{E}\mathcal{F}_{\mathscr{X}}(\mathbf{\Theta_0})| \ge (1-\kappa)d^{-2}\|\mathbf{\Theta_0}\|_{\rm F}^2 + T_0\}. \tag{C.5}$$

Let $D_0 = (\alpha^* d)^2(\psi\delta\log(2d)/n)^{1/2}$, then by (4.14) we have $\|\mathbf{\Theta_0}\|_{\rm F}^2 \ge D_0$, so by a specific $\beta > 1$ (that will be selected later), there exists positive integer l such that $\|\mathbf{\Theta_0}\|_{\rm F}^2 \in [\beta^{l-1}D_0, \beta^l D_0)$. We further consider $\mathcal{C}(\psi, l) = \mathcal{C}(\psi)\cap\{\mathbf{\Theta} : \|\mathbf{\Theta}\|_{\rm F}^2 \in [\beta^{l-1}D_0, \beta^l D_0)\}$, and define a term

$$\mathcal{Z}_{\mathscr{X}}(l) = \sup_{\mathbf{\Theta}\in\mathcal{C}(\psi,l)} |\mathcal{F}_{\mathscr{X}}(\mathbf{\Theta}) - \mathbb{E}\mathcal{F}_{\mathscr{X}}(\mathbf{\Theta})|,$$

then we know the event defined in (C.5) implies the event

$$\mathscr{B}_l = \{\mathcal{Z}_{\mathscr{X}}(l) \ge (1-\kappa)d^{-2}\beta^{l-1}D_0 + T_0\}. \tag{C.6}$$

By taking the union bound over $l \in \mathbb{N}^*$ we obtain $\mathbb{P}(\mathscr{B}) \le \sum_{l=1}^\infty \mathbb{P}(\mathscr{B}_l)$.
II. It suffices to bound $\mathbb{P}(\mathscr{B}_l)$. We first bound the deviation $|\mathcal{Z}_{\mathscr{X}}(l) - \mathbb{E}\mathcal{Z}_{\mathscr{X}}(l)|$. We consider $\widetilde{\mathscr{X}} = (\widetilde{\boldsymbol{X}_1}, \boldsymbol{X_2}, ..., \boldsymbol{X_n})$ where only the first component may be different from $\mathscr{X}$

$$\begin{aligned}
&\sup_{\mathscr{X},\widetilde{\mathscr{X}}} |\mathcal{Z}_{\mathscr{X}}(l) - \mathcal{Z}_{\widetilde{\mathscr{X}}}(l)| = \sup_{\mathscr{X},\widetilde{\mathscr{X}}}\Big|\sup_{\mathbf{\Theta}\in\mathcal{C}(\psi,l)}|\mathcal{F}_{\mathscr{X}}(\mathbf{\Theta}) - \mathbb{E}\mathcal{F}_{\mathscr{X}}(\mathbf{\Theta})| - \sup_{\mathbf{\Theta}\in\mathcal{C}(\psi,l)}|\mathcal{F}_{\widetilde{\mathscr{X}}}(\mathbf{\Theta}) - \mathbb{E}\mathcal{F}_{\widetilde{\mathscr{X}}}(\mathbf{\Theta})|\Big| \\
&\le \sup_{\mathscr{X},\widetilde{\mathscr{X}}}\Big|\sup_{\mathbf{\Theta}\in\mathcal{C}(\psi,l)}|\mathcal{F}_{\mathscr{X}}(\mathbf{\Theta}) - \mathcal{F}_{\widetilde{\mathscr{X}}}(\mathbf{\Theta})|\Big| = \sup_{\boldsymbol{X_1},\widetilde{\boldsymbol{X_1}}}\sup_{\mathbf{\Theta}\in\mathcal{C}(\psi,l)}\frac{1}{n}\Big|\big\langle\boldsymbol{X_1},\mathbf{\Theta}\big\rangle|^2 - |\big\langle\widetilde{\boldsymbol{X_1}},\mathbf{\Theta}\big\rangle|^2\Big| \le \frac{4(\alpha^*)^2}{n}.
\end{aligned}$$

Note that n components of $\mathscr{X}$ are symmetrical, by bounded different inequality (e.g., Corollary 2.21, [88]), for any $t > 0$ we have

$$\mathbb{P}\Big(\mathcal{Z}_{\mathscr{X}}(l) - \mathbb{E}\mathcal{Z}_{\mathscr{X}}(l) \ge t\Big) \le \exp\Big(-\frac{nt^2}{8(\alpha^*)^4}\Big). \tag{C.7}$$

It remains to bound $\mathbb{E}\mathcal{Z}_{\mathscr{X}}(l)$. Let $\mathscr{E} = (\varepsilon_1, ..., \varepsilon_n)$ be i.i.d. Rademacher random variables satisfying $\mathbb{P}(\varepsilon_k = 1) = \mathbb{P}(\varepsilon_k = -1) = 1/2$, then by symmetrization of expectations (e.g., Theorem 16.1, [85]), Talagrand's inequality (e.g., Theorem 16.2, [85]), the second constraint in (4.14), it yields that

$$\begin{aligned}
&\mathbb{E}\mathcal{Z}_{\mathscr{X}}(l) = \mathbb{E}\sup_{\mathbf{\Theta}\in\mathcal{C}(\psi,l)}|\mathcal{F}_{\mathscr{X}}(\mathbf{\Theta}) - \mathbb{E}\mathcal{F}_{\mathscr{X}}(\mathbf{\Theta})| = \mathbb{E}\sup_{\mathbf{\Theta}\in\mathcal{C}(\psi,l)}\Big|\frac{1}{n}\sum_{k=1}^n\big\{\big\langle\boldsymbol{X_k},\mathbf{\Theta}\big\rangle^2 - \mathbb{E}\big\langle\boldsymbol{X_k},\mathbf{\Theta}\big\rangle^2\big\}\Big| \\
&\le 2\mathbb{E}_{\mathscr{X}}\mathbb{E}_{\mathscr{E}}\sup_{\mathbf{\Theta}\in\mathcal{C}(\psi,l)}\Big|\frac{1}{n}\sum_{k=1}^n\varepsilon_k\big\langle\boldsymbol{X_k},\mathbf{\Theta}\big\rangle^2\Big| \le 16\alpha^*\mathbb{E}\sup_{\mathbf{\Theta}\in\mathcal{C}(\psi,l)}\Big|\big\langle\frac{1}{n}\sum_{k=1}^n\varepsilon_k\boldsymbol{X_k},\mathbf{\Theta}\big\rangle\Big| \\
&\le 16\alpha^*\mathbb{E}\Big\|\frac{1}{n}\sum_{k=1}^n\varepsilon_k\boldsymbol{X_k}\Big\|_{\rm op}\sup_{\mathbf{\Theta}\in\mathcal{C}(\psi,l)}\|\mathbf{\Theta}\|_{\rm nu} \le 160\alpha^* r^{\frac{1}{2-q}}\{\beta^l D_0\}^{\frac{1-q}{2-q}}\mathbb{E}\Big\|\frac{1}{n}\sum_{k=1}^n\varepsilon_k\boldsymbol{X_k}\Big\|_{\rm op}.
\end{aligned} \tag{C.8}$$

Assume $d\log(2d)/n < 1/16$, by matrix bernstein inequality (Theorem 6.1.1, [84]) it holds that

$\mathbb{E}\Big\|\frac{1}{n}\sum_{k=1}^{n}\varepsilon_k\boldsymbol{X_k}\Big\|_{\text{op}}\le\frac{3}{2}\sqrt{\frac{\log(2d)}{nd}}$. We then plug it in (C.8), some algebra yields

$$\mathbb{E}\mathcal{Z}_{\mathscr{X}}(l)\le\big\{(2-q)T_0\big\}^{\frac{1}{2-q}}\big\{d^{-2}\beta^lD_0\big\}^{\frac{1-q}{2-q}}\le\frac{1-q}{2-q}\frac{\beta^lD_0}{d^2}+T_0 \tag{C.9}$$

By combining with (C.6), (C.7) and let $\kappa_1=\frac{1-\kappa}{\beta}-\frac{1-q}{2-q}$ (here we assume $\kappa_1\in(0,1)$ since we can choose κ sufficiently close to 0, β sufficiently close to 1), we have

$$\mathbb{P}(\mathscr{B}_l)\le\mathbb{P}\Big(\mathcal{Z}_{\mathscr{X}}(l)-\mathbb{E}\mathcal{Z}_{\mathscr{X}}(l)\ge\kappa_1\frac{\beta^lD_0}{d^2}\Big)\le\exp\Big(-\frac{n\kappa_1^2\beta^{2l}D_0^2}{8(\alpha^*d)^4}\Big). \tag{C.10}$$

We further plug in D_0 and use $\beta^{2l}\ge 2l\log\beta$, it yields that

$$\mathbb{P}(\mathscr{B})\le\sum_{l=1}^{\infty}\mathbb{P}(\mathscr{B}_l)\le\sum_{l=1}^{\infty}\big[(2d)^{-\frac{\psi\delta\kappa_1^2\log\beta}{4}}\big]^l\le d^{-\delta},$$

the last inequality holds since we can let ψ be large such that $\psi\ge4(\kappa_1^2\log\beta)^{-1}$. □

Proof of Theorem 11. **I.** By Lemma 3 we can choose sufficiently large C_{14} in (4.17) to ensure (4.7) holds with probability higher than $1-2d^{1-\delta}$, then (4.8) holds with high probability. From Lemma 4 we can further rule out probability $d^{-\delta}$ to ensure (4.15) holds.

By (4.5) and (4.6) we have $\|\widehat{\boldsymbol{\Delta}}\|_{\max}\le\|\widehat{\boldsymbol{\Theta}}\|_{\max}+\|\boldsymbol{\Theta^*}\|_{\max}\le2\alpha^*$. Thus, the estimation error satisfies the first constraint of $\mathcal{C}(\psi)$. Since (4.8) displays the second constraint in $\mathcal{C}(\psi)$, whether $\widehat{\boldsymbol{\Delta}}\in\mathcal{C}(\psi)\in\mathcal{C}(\psi)$ holds only depends on the third constraint, and let us discuss as follows:

Case 1. $\widehat{\boldsymbol{\Delta}}\notin\mathcal{C}(\psi)$. Note that it can only violate the third constraint of $\mathcal{C}(\psi)$, so we know that $\|\widehat{\boldsymbol{\Delta}}\|_{\text{F}}^2\le(\alpha^*d)^2\sqrt{\frac{\psi\delta\log(2d)}{n}}$. Under the assumption $r\gtrsim d^q$, $n\lesssim d^2\log(2d)$, it holds that

$$\frac{\|\widehat{\boldsymbol{\Delta}}\|_{\text{F}}^2}{d^2}\lesssim(\alpha^*)^2\sqrt{\frac{\delta\log(2d)}{n}}\lesssim(\alpha^*)^2\frac{\delta d\log d}{n}\lesssim rd^{-q}\Big((\alpha^*)^2\frac{\delta d\log d}{n}\Big)^{1-q/2}. \tag{C.11}$$

Case 2. $\widehat{\boldsymbol{\Delta}}\in\mathcal{C}(\psi)$. By (4.15) we know $\mathcal{F}_{\mathscr{X}}(\widehat{\boldsymbol{\Delta}})\ge\kappa d^{-2}\|\widehat{\boldsymbol{\Delta}}\|_{\text{F}}^2-T_0$. If $T_0\ge\frac{1}{2}\kappa d^{-2}\|\widehat{\boldsymbol{\Delta}}\|_{\text{F}}^2$, then we plug in T_0 and obtain

$$\frac{\|\widehat{\boldsymbol{\Delta}}\|_{\text{F}}^2}{d^2}\lesssim rd^{-q}\Big((\alpha^*)^2\frac{\delta d\log d}{n}\Big)^{1-q/2}. \tag{C.12}$$

If $T_0\le\frac{1}{2}\kappa d^{-2}\|\widehat{\boldsymbol{\Delta}}\|_{\text{F}}^2$, then we have $\mathcal{F}_{\mathscr{X}}(\widehat{\boldsymbol{\Delta}})\ge\frac{1}{2}\kappa d^{-2}\|\widehat{\boldsymbol{\Delta}}\|_{\text{F}}^2$. Note that this displays the RSC in (4.9), so we now use Corollary 3 and obtain

$$\frac{\|\widehat{\boldsymbol{\Delta}}\|_{\text{F}}^2}{d^2}\lesssim rd^{-q}\Big(\max\{(\alpha^*)^2,\sigma^2\}\log d\log n\frac{\delta d}{n}\Big)^{1-q/2}. \tag{C.13}$$

Now we can see that in all cases considered above, the bound of $\|\widehat{\boldsymbol{\Delta}}\|_{\text{F}}^2/d^2$ in (4.18) holds. Then a direct application of (4.8) delivers the bound of $\|\widehat{\boldsymbol{\Delta}}\|_{\text{nu}}/d$ in (4.18). To conclude, (4.18) holds with probability higher than $1-3d^{1-\delta}$. □

C.2 Heavy-tailed Data

Proof of Lemma 5. From (4.1) we have $\mathbb{E}(Y_k \boldsymbol{X_k}) = \mathbb{E}(\langle \boldsymbol{X_k}, \boldsymbol{\Theta}^* \rangle \boldsymbol{X_k})$, and since $\eta < \gamma$ by Lemma 1 we know $\mathbb{E}_{\Lambda_k}(\gamma \cdot \dot{Y}_k) = \widetilde{Y}_k$, hence we have

$$\Big\|\frac{1}{n}\sum_{k=1}^n \Big[\langle \boldsymbol{X_k}, \boldsymbol{\Theta}^* \rangle - \gamma \cdot \dot{Y}_k\Big]\boldsymbol{X_k}\Big\|_{\text{op}} \leq \Big\|\frac{1}{n}\sum_{k=1}^n \big[\gamma \cdot \dot{Y}_k \boldsymbol{X_k} - \mathbb{E}(\gamma \cdot \dot{Y}_k \boldsymbol{X_k})\big]\Big\|_{\text{op}} +$$

$$\Big\|\mathbb{E}\big[(\widetilde{Y}_k - Y_k)\boldsymbol{X_k}\big]\Big\|_{\text{op}} + \Big\|\frac{1}{n}\sum_{k=1}^n \big[\langle \boldsymbol{X_k}, \boldsymbol{\Theta}^* \rangle \boldsymbol{X_k} - \mathbb{E}(\langle \boldsymbol{X_k}, \boldsymbol{\Theta}^* \rangle \boldsymbol{X_k})\big]\Big\|_{\text{op}} := R_1 + R_2 + R_3.$$

<u>*Bound of R_1, R_3.*</u> We use matrix Bernstein inequality (Theorem 6.1.1, [84]), and the arguments are exactly the same as the corresponding parts in the proof of Lemma 3. As a result, one can still invoke Matrix Bernstein to show (C.3) and (C.4), but only with different value of γ. To obtain the explicit form of the bounds, we further plug in γ in (4.20), with probability higher than $1 - 2d^{1-\delta}$ it gives

$$\max\{R_1, R_3\} \lesssim \max\{\alpha^*, \sqrt{M}\}\left(\frac{\delta \log d}{nd^3}\right)^{1/4}.$$

<u>*Bound of R_2.*</u> Let $X_{k,ij}$ be the (i,j)-th entry of $\boldsymbol{X_k}$, where $(i,j) \in [d] \times [d]$ is fixed, we first bound the element-wise error $|\mathbb{E}(\widetilde{Y}_k - Y_k)X_{k,ij}|$. Recall the definition of truncation, Lemma 1 gives

$$|\mathbb{E}\big[(\widetilde{Y}_k - Y_k)X_{k,ij}\big]| = |\mathbb{E}\big[(\widetilde{Y}_k - Y_k)X_{k,ij}\mathbb{1}(|Y_k| > \eta)\big]| \leq \mathbb{E}|Y_k|X_{k,ij}\mathbb{1}(|Y_k| > \eta).$$

Note that $X_{k,ij}$ can only be 1 or 0, and $\mathbb{P}(X_{k,ij} = 1) = d^{-2}$. Let Θ^*_{ij} be the (i,j)-th entry of $\boldsymbol{\Theta}^*$, we further compute it via law of total expectation, then use Cauchy-Schwarz inequality and Marcov's inequality, finally plug in η finally. These steps deliver

$$\begin{aligned}
&\mathbb{E}|Y_k|X_{k,ij}\mathbb{1}(|Y_k| > \eta) = \mathbb{E}\Big(\mathbb{E}\Big[|Y_k|X_{k,ij}\mathbb{1}(|Y_k| > \eta)\Big|X_{k,ij}\Big]\Big) \\
&= d^{-2}\mathbb{E}\Big[|Y_k|X_{k,ij}\mathbb{1}(|Y_k| > \eta)\Big|X_{k,ij} = 1\Big] = d^{-2}\mathbb{E}\Big[|\Theta^*_{ij} + \epsilon_k|\mathbb{1}(|\Theta^*_{ij} + \epsilon_k| \geq \eta)\Big] \\
&\leq d^{-2}\sqrt{\mathbb{E}|\Theta^*_{ij} + \epsilon_k|^2\mathbb{P}(|\Theta^*_{ij} + \epsilon_k| \geq \eta)} \leq d^{-2}\eta^{-1}\mathbb{E}|\Theta^*_{ij} + \epsilon_k|^2 \\
&\leq 2d^{-2}\eta^{-1}[(\Theta^*_{ij})^2 + \mathbb{E}\epsilon_k^2] \leq 4(dC_{15})^{-1}\max\{\alpha^*, \sqrt{M}\}\Big(\frac{\delta \log d}{nd^3}\Big)^{1/4}
\end{aligned}$$

Since the above analysis works for all $(i,j) \in [d] \times [d]$, this is also an upper bound for the max norm, which delivers a bound for operator norm

$$R_2 \leq d\Big\|\mathbb{E}\big[(\widetilde{Y}_k - Y_k)\boldsymbol{X_k}\big]\Big\|_{\max} \lesssim \max\{\alpha^*, \sqrt{M}\}\Big(\frac{\delta \log d}{nd^3}\Big)^{1/4}. \tag{C.14}$$

The result follows from the upper bounds for R_1, R_2, R_3. □

Proof of Theorem 12. By Lemma 5 we can choose sufficiently large C_{17} in (4.22) to ensure (4.7) holds with probability higher than $1 - 2d^{1-\delta}$, then it further implies (4.8), meaning that $\widehat{\boldsymbol{\Delta}}$ satisfies the second constraint of $\mathcal{C}(\psi)$. From Lemma 4 we can further rule out probability

$d^{-\delta}$ so that (4.15) holds. Evidently we have $\|\widehat{\boldsymbol{\Delta}}\|_{\max} \le 2\alpha^*$. Thus, with probability higher than $1-3d^{1-\delta}$, $\widehat{\boldsymbol{\Delta}}$ satisfies the first two constraints of $\mathcal{C}(\psi)$, and (4.15) holds. Based on these conditions we further discuss as follows:

Case 1. $\widehat{\boldsymbol{\Delta}} \notin \mathcal{C}(\psi)$, then by exactly the same analysis in proof of Theorem 11, we obtain (C.11).

Case 2. $\widehat{\boldsymbol{\Delta}} \in \mathcal{C}(\psi)$, then we have $\mathcal{F}_{\mathscr{X}}(\widehat{\boldsymbol{\Delta}}) \ge \kappa d^{-2}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2 - T_0$. If $T_0 \ge \frac{1}{2}\kappa d^{-2}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2$, then (C.12) holds. Otherwise, we have the restricted strong convexity $\mathcal{F}_{\mathscr{X}}(\widehat{\boldsymbol{\Delta}}) \ge \frac{1}{2}\kappa d^{-2}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2$. We then apply Corollary 3 and plug in λ, it holds that

$$\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2/d^2 \lesssim rd^{-q}\big(\max\{(\alpha^*)^2, M\}\sqrt{\frac{\delta d\log d}{n}}\big)^{1-q/2}. \tag{C.15}$$

It is not hard to see that the right hand side of (C.15) dominates the bound in (C.11) and (C.12), so the bound for $\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2/d^2$ in (4.23) holds. The bound for $\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{nu}}/d$ follows from a direct application of (4.8). □

D Comparisons with Related Work

D.1 1-bit Compressed Sensing

In this part we compare our Theorem 9, 10 with existing results of 1-bit CS. Among improvements of different perspectives, we would emphasize our contributions of extending 1-bit CS to non-Gaussian sensing vectors, and the first tractable recovery convex programming recovery method for heavy-tailed sensing vectors.

The traditional setting of 1-bit CS, where one needs to recovery a sparse d-dimensional signal Θ^* based on measurement $\dot{Y}_k = \mathrm{sign}(X_k^T\Theta^*)$ with some X_k, was first introduced in [15] and widely studied in subsequent works (e.g., [50,69,70]). By projection-based method [72] or K-Lasso [71], similar results have been obtained for a model with more general observation and signal assumption. Nevertheless, all these results are restricted to Gaussian sensing vectors that can be unrealistic in practice[3]. There does exist one work, [1], presents result for X_k with i.i.d. sub-Gaussian entries. However, the result in [1] is overly restrictive and impractical, see the discussions in [38] for more details.

To overcome the Gaussian restriction (and also some other limitations), recent works show introducing dithering noise can go a long way. With dithering noise Λ_k, the measurement now becomes $\dot{Y}_k = \mathrm{sign}(X_k^T\Theta^* + \Lambda_k)$. In this setting, specifically, we can recover the signal with norm information [58], achieve exponentially-decaying error rate (This requires adaptive dithering) [5], and perhaps more prominently, accommodate non-Gaussian X_k [37,38,83]. In what follows, we will focus on comparing Theorem 9, 10 with the most related works [38,83] that adopt uniform dithering noise (by contrast, [5,58] use Gaussian dithering noise). The comparisons will be conducted on exactly sparse Θ^* since [38,83] do not adopt the formulation $\sum_{k=1}^n |\theta_k^*|^q \le s,\ q\in(0,1)$ for approximately sparse Θ^*. For other developments on 1-bit CS (or more generally, quantized compressed sensing), we refer readers to the survey papers [16,35].

Dirksen and Mendelson [38] first essentially extend 1-bit CS to non-Gaussian X_k. Their methodology is aligned with classic works like [70], that is, to start from the viewpoint of

[3] More precisely, [71] handles $X_k \sim \mathcal{N}(0,\boldsymbol{\Sigma})$ with unknown $\boldsymbol{\Sigma}$ while other several papers above assume $X_k \sim \mathcal{N}(0,\boldsymbol{I_d})$.

random hyperplane tessellation. Specifically, under sub-Gaussian or even heavy-tailed X_k[4] with uniform dithering noise, they show a relatively small number of random hyperplanes (that depends on the complexity of Θ^*) leads to ρ-uniform tessellation on the signal set of interest. Moreover, they apply the new hyperplane tessellation results to 1-bit CS and propose two reconstruction optimization problems

$$\begin{cases} \text{(a)}: \widehat{\Theta} \in \arg\min_{\Theta \in \mathbb{R}^d} \sum_{k=1}^{n} \mathbb{1}(\text{sign}(X_k^T\Theta + \Lambda_k) \neq Y_k), \quad \text{s.t. } \|\Theta\|_0 \leq s, \ \|\Theta\|_2 \leq 1 \\ \text{(b)}: \widehat{\Theta} \in \arg\min_{\Theta \in \mathbb{R}^d} \frac{1}{2\lambda}\|\Theta\|_2^2 - \frac{1}{2n}\sum_{k=1}^{n} Y_k X_k^T \Theta, \quad \text{s.t. } \|\Theta\|_1 \leq s, \ \|\Theta\|_2 \leq 1 \end{cases}. \tag{D.1}$$

Although (a) is shown to possess uniform recovery guarantee with fast rate in both sub-Gaussian and heavy-tailed X_k, it is essentially intractable due to the ℓ_0 constraint and the 0-1 objective function. Also, a secondary drawback is that, the information of Λ_k is needed in problem (a), which induces undesired memory or transmission costs. For these reasons, (a) is mainly of theoretical interest. To address the issue, (b) is proposed as a convex relaxation of (a). Under sub-Gaussian X_k and ϵ_k, the error rate of (b) for s-sparse Θ^* was shown to be $\tilde{O}\big(\sqrt[4]{\frac{s}{n}}\big)$, and this is significantly inferior to the near optimal rate $\tilde{O}\big(\sqrt{\frac{s}{n}}\big)$ delivered by our Theorem 9. Indeed, their error rate for sub-Gaussian data is even worse than our rate for heavy-tailed data, i.e., $\tilde{O}\big(\sqrt[3]{\frac{s^2}{n}}\big)$ given in Theorem 10, while the guarantee of (b) under heavy-tailed X_k has not yet been established.

Besides the error rate, all results in [38] assume X_k is isotropic, i.e., $\mathbf{\Sigma}_{XX} = \boldsymbol{I_d}$. By contrast, our theory can handle unknown $\mathbf{\Sigma}_{XX}$ satisfying $\lambda_{\min}(\mathbf{\Sigma}_{XX}) = \Omega(1)$. Actually, since Assumption 3 nicely encompasses isotropic X_k, we can even carry out the more tricky 1-bit QC-CS under their conditions, see Theorem 7, 8 for the setting with known $\mathbf{\Sigma}_{XX}$.

We further give two side remarks to end the comparison with [38]. Firstly, it is necessary to have pre-estimation of $\|\Theta\|_1$ to invoke (b), while this is avoided in our unconstrained recovery program. Secondly, their results are only for symmetric X_k, while we only assume zero mean. On the other hand, the advantages of [38] due to Dirksen and Mendelson may be more general signal set, lower moment requirement in heavy-tailed X_k, and a partial extension to structured random measurement matrix in the companion work [37]. These are left as future research directions of our theories.

A result directly comparable to our Theorem 9 is due to Thrampoulidis and Rawat [83]. Under almost the same setting they assume $\Theta^* \in T$ and consider a constrained Lasso

$$\widehat{\Theta} \in \arg\min_{\Theta \in T} \frac{1}{2n}\sum_{k=1}^{n} \big(X_k^T\Theta - \gamma \cdot \dot{Y}_k\big)^2. \tag{D.2}$$

This is analogous to our convex programming problem (3.24): Up to constant, the objective of (D.2) equals the loss function (i.e., the first two terms) in (3.24); And the only difference is that, the structure of Θ^*, specifically sparsity, is incorporated into (D.2) via the constraint, but appears in (3.24) as a regularizer. The recovery guarantee is given in Theorem IV.1 in [83]. Interestingly, when restricted to exactly s-sparse Θ^*, they choose $T = \{\Theta : \|\Theta\|_1 \leq \|\Theta^*\|_1\}$ and show ℓ_2 norm error rate $\tilde{O}\big(\sqrt{\frac{s}{n}}\big)$ that coincides with Theorem 9. Despite these similarities,

[4]In [38], heavy-tailed X_k is assumed to satisfy $\mathbb{E}\big(|v^TX_k|^2\big) \leq L\big(\mathbb{E}|v^TX_k|\big)^2$ for any $v \in \mathbb{R}^d$.

our Theorem 9 exhibits several obvious improvements. Firstly, we consider pre-quantization noise ϵ_k while they only study noiseless case. Secondly, we assume zero-mean X_k satisfies $\lambda_{\min}(\mathbf{\Sigma}_{XX}) = \Omega(1)$, but [83] requires symmetric X_k to satisfy a nondegeneracy condition formulated as $\inf_{\|v\|_2=1} \mathbb{E}|v^T X_k| = \Omega(1)$, which is more restrictive. Thirdly, their guarantee is valid with probability at least 0.99, while our probability term $1-O(d^{2-\delta})$ is finer. In addition, we comment that a pre-estimation of $\|\Theta^*\|_1$ is needed to specify T and invoke (D.2), while our unconstrained program (3.24) is free of this issue and hence is more practically appealing. Thus, even without mentioning our result for 1-bit CS in heavy-tailed regime, Theorem 9 can represent the all-round improvement of Theorem IV.1 in [83]. Indeed, their only advantage seems to be the more general assumption $\Theta^* \in T$, while we believe our result straightforwardly extends to other interesting signal structures such as (approximate) low-rankness.

In [5, 58] the authors study $\dot{Y}_k = \text{sign}(X_k^T \Theta^* + \Lambda_k)$ where Λ_k is Gaussian dithering noise. Convex programming problems are also proposed in both papers to recover s-sparse Θ^*, but X_k is restricted to standard Gaussian sensing vector. Specifically, the theoretical rate in Theorem 4 of [58] reads $\tilde{O}\big(\sqrt[5]{\frac{s}{n}}\big)$, and Theorem 2 of [5] gives the guarantee $\tilde{O}\big(\sqrt[4]{\frac{s}{n}}\big)$. This is obviously slower than the rates presented in our Theorem 9 (sub-Gaussian X_k), Theorem 10 (heavy-tailed X_k).

In a nutshell, this work gains significant improvements on current results of 1-bit CS, in terms of generality of sensing vectors, faster rate and computational feasibility.

D.2 1-bit Matrix Completion

In this part, we compare our Theorem 11, 12 with existing results on 1-bit matrix completion (1-bit MC), a problem first proposed and studied in [23, 32]. We shall see that, our work presents the first result in 1-bit MC that can handle pre-quantization random noise with unknown distribution.

Unlike 1-bit CS where one can still recover the direction $\Theta^*/\|\Theta^*\|_2$ from the direct quantized measurement $\dot{Y}_k = \text{sign}(X_k^T \Theta^*)$, due to the nature of the covariate in matrix completion (i.e., $\boldsymbol{X_k} = e_{i(k)} e_{j(k)}^T$), 1-bit MC could be extremely ill-posed if we only observe $\dot{Y}_k = \text{sign}\big(\langle \boldsymbol{X_k}, \boldsymbol{\Theta}^* \rangle\big)$, even when $\boldsymbol{\Theta}^*$ is rank-1, see the discussion in [32]. Thus, to proceed the study of 1-bit MC, dithering noise (denoted by Λ_k) is indispensable for the well-posedness, hence the measurement (i.e., observation) becomes

$$\dot{Y}_k = \text{sign}\big(\langle \boldsymbol{X_k}, \boldsymbol{\Theta}^* \rangle + \Lambda_k\big).$$

Existing works consider dithering noise with rather general distribution, but mainly emphasize Logistic model and Probit model[5].

Let us give a brief review of existing results. In the first study of 1-bit MC [32], Davenport et al. proposed to recover $\boldsymbol{\Theta}^*$ via negative log-likelihood minimization (put $d_1 = d_2 = d$)

$$\widehat{\boldsymbol{\Theta}} \in \mathop{\arg\min}_{\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}} \ \mathcal{L}_{\text{NLL}}(\boldsymbol{\Theta}), \quad \text{s.t. } \|\boldsymbol{\Theta}\|_{\max} \le \alpha^*, \ \|\boldsymbol{\Theta}\|_{\text{nu}} \le \alpha^* d\sqrt{r}\ . \tag{D.3}$$

In (D.3), the first constraint is commonly used in matrix completion (see the interpretation at the beginning of Section 4), the second constraint relaxes $\text{rank}(\boldsymbol{\Theta}) \le r$ via the relation

$$\|\boldsymbol{\Theta}\|_{\text{nu}} \le \sqrt{\text{rank}(\boldsymbol{\Theta})}\|\boldsymbol{\Theta}\|_{\text{F}} \le d\|\boldsymbol{\Theta}\|_{\max}\sqrt{\text{rank}(\boldsymbol{\Theta})},$$

[5]The Probit model corresponds to Gaussian dithering noise $\Lambda_k \sim \mathcal{N}(0, \sigma^2)$.

and the loss function $\mathcal{L}_{\text{NLL}}(\mathbf{\Theta})$ is

$$\begin{aligned}\mathcal{L}_{\text{NLL}}(\mathbf{\Theta}) = -\frac{1}{n}\sum_{k=1}^{n}\Big[&\mathbb{1}(\dot{Y}_k = 1)\log\mathbb{P}(\big\langle \boldsymbol{X_k}, \mathbf{\Theta}\big\rangle + \Lambda_k \geq 0) \\ &+ \mathbb{1}(\dot{Y}_k = -1)\log\mathbb{P}(\big\langle \boldsymbol{X_k}, \mathbf{\Theta}\big\rangle + \Lambda_k < 0)\Big].\end{aligned} \tag{D.4}$$

Some developments can be seen in subsequent works, to name a few, [23] used another surrogate of matrix rank rather than the nuclear norm[6], [10, 56, 62] extended 1-bit MC to finite alphabets, [9,10,68] imposed (exactly) low-rank constraint without relaxation, [56,62] adopted nuclear norm penalty to avoid the pre-estimation of $\|\mathbf{\Theta}^*\|_{\text{nu}}$ needed in (D.3).

We stress that all above works are restricted to a noiseless setting, by saying this, we do not regard Λ_k as a detrimental noise since the dithering is indeed beneficial to the recovery. When it comes to noise that may corrupt the recovery, we are aware of only two recent papers [46,80]. In [46], Gao et al. considered a deterministic sparse pattern $\boldsymbol{S}^*$ mixing with the desired low-rank structure $\mathbf{\Theta}^*$. Specifically, this more general "low-rank plus sparse" model (in one-bit setting) can be formulated as

$$\dot{Y}_k = \text{sign}\big(\big\langle \boldsymbol{X_k}, \mathbf{\Theta}^* + \boldsymbol{S}^*\big\rangle + \Lambda_k\big), \quad \text{where } \|\text{vec}(\boldsymbol{S}^*)\|_0 \leq s. \tag{D.5}$$

In [80], Shen et al. studied 1-bit MC with post-quantization noise in a form of sign flipping, which can be described by

$$\dot{Y}_k = \delta_k \cdot \text{sign}\big(\big\langle \boldsymbol{X_k}, \mathbf{\Theta}^*\big\rangle + \Lambda_k\big), \quad \text{where } \mathbb{P}(\delta_k = -1) = \tau_0,\ \mathbb{P}(\delta_k = 1) = 1 - \tau_0. \tag{D.6}$$

Evidently, for (D.5) or (D.6), as done in [46, 80], the recovery can still be based on negative log-likelihood minimization. However, if we consider pre-quantization noise ϵ_k with unknown distribution (This is a natural and well-studied situation in other statistical estimation models), i.e.,

$$\dot{Y}_k = \text{sign}\big(\big\langle \boldsymbol{X_k}, \mathbf{\Theta}^*\big\rangle + \Lambda_k + \epsilon_k\big), \tag{D.7}$$

recovery based on likelihood no longer works due to lack of knowledge of $\mathcal{L}_{\text{NLL}}(\mathbf{\Theta})$. Therefore, before our work, it was an open question whether 1-bit MC under unknown pre-quantization random noise is possible.

Our Theorem 11, 12 affirmatively answer this open question. Particularly, under uniformly distributed Λ_k, sub-Gaussian or even heavy-tailed ϵ_k, we formulate 1-bit CS as a convex programming problem and establish theoretical guarantee. This is due to a novel loss function deviating from existing papers, that is, we now use a generalized quadratic loss (see (4.6))

$$\mathcal{L}(\mathbf{\Theta}) = \frac{1}{2n}\sum_{k=1}^{n}\big(\big\langle \boldsymbol{X_k}, \mathbf{\Theta}\big\rangle - \gamma\cdot\dot{Y}_k\big)^2. \tag{D.8}$$

To see the essential difference, one may compare (D.8) and the explicit form of (D.4) when $\Lambda_k \sim \text{uni}\big([-\gamma,\gamma]\big)$. For the core idea behind, while maximum likelihood estimation is a quite standard estimation strategy, the inspiration of (D.8) is drawn from Lemma 1, i.e., $\gamma\cdot\dot{Y}_k$ can serve as a surrogate of the full observation $Y_k = \big\langle \boldsymbol{X_k}, \mathbf{\Theta}^*\big\rangle + \epsilon_k$.

[6]This rank surrogate is called max-norm but totally different from $\|.\|_{\max}$ in our work. To avoid confusion,

At first glance, one may feel that (D.8) is a bit coarse compared with negative log-likelihood, but it is striking that under sub-Gaussian noise our estimator achieves near minimax rate (Theorem 11). For comparison, we go back to the noiseless (i.e., $\epsilon_k = 0$ in (D.7)) and exactly low-rank (i.e., $q = 0$ in (4.3)) case. In this case, Theorem 11 gives a bound $\tilde{O}\big((\alpha^*)^2 \frac{rd}{n}\big)$ for mean squared error. This is faster than $\tilde{O}\big((\alpha^*)^2 \sqrt{\frac{rd}{n}}\big)$ obtained in two pioneering works [23,32], and similar to the more recent paper [62].

To sum up, we present the first result for 1-bit MC with unknown pre-quantization random noise, which can either be sub-Gaussian or heavy-tailed. In addition, by some extra technicalities, we believe our method can be extended to both deterministic sparse corruption in [46] and sign flipping noise in [80]. On the other hand, our restriction is that Λ_k should be uniformly distributed, and it would be an interesting open question whether our method can be extended to other dithers.

E Details and Algorithms in Experiments

E.1 Sparse Covariance Matrix Estimation

Simulation Details. To generate the $d \times d$ underlying covariance matrix $\mathbf{\Sigma^*}$ that satisfies Assumption 1 with $q = 0$ and sparsity s, we first construct

$$\mathbf{\Sigma_0^*} = \begin{pmatrix} \mathbf{\Sigma_1^*} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I_{d-3s}} \end{pmatrix},$$

where $\mathbf{\Sigma_1^*} = \mathrm{diag}(\mathbf{\Sigma_2^*}, \mathbf{\Sigma_2^*}, \mathbf{\Sigma_2^*}) \in \mathbb{R}^{3s\times 3s}$, and $\mathbf{\Sigma_2^*} = [\sigma^*_{2,ij}] \in \mathbb{R}^{s\times s}$ are defined as $\sigma^*_{2,ii} = 1$ for $i \in [s]$, $\sigma^*_{2,12} = \sigma^*_{2,21} = 0.99 - (s-2)\cdot 0.03$, $\sigma^*_{2,ij} = 0.03$ for all other entries. By normalizing the operator norm, we set

$$\mathbf{\Sigma^*} = \frac{\mathbf{\Sigma_0^*}}{\|\mathbf{\Sigma_0^*}\|_{\mathrm{op}}}.$$

We i.i.d. draw sub-Gaussian $X_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma^*})$, and draw heavy-tailed X_k from Student's t distribution via the Matlab function "mvtrnd(·)" with $\nu = 6$. Then, we apply the one-bit quantization scheme with parameters slightly tuned to be well-functioning, to obtain the binary data $\big\{\dot{X}_{kj} : k \in [n], j = 1, 2\big\}$. Now, we can directly construct the 1-bit estimator $\widehat{\mathbf{\Sigma}}$ defined in (2.4), (2.18), and track the experimental recovery error. In our results, each experiment is obtained as the mean value of 15 independent runs.

E.2 Sparse Linear Regression

Simulation Details. We conduct numerical experiments of 1-bit QC-CS (Theorem 7, 8) and 1-bit CS (Theorem 9, 10). We consider isotropic covariate (i.e., $\mathbb{E}X_kX_k^T = \boldsymbol{I_d}$) that is kind of convention in compressed sensing (e.g., [25, 72]), which admits Assumption 3 required for 1-bit QC-CS. For the covariate, specifically, sub-Gaussian X_k are generated from Gaussian distribution, while entries of heavy-tailed X_k are i.i.d. drawn from $\sqrt{\frac{2}{3}} \cdot t(\nu = 6)$. Here, $t(\nu = 6)$ represents Student's t distribution with 6 degrees of freedom, and $\sqrt{\frac{2}{3}}$ aims to

we refer readers to [23] for the details.

normalize the variance. We set the first s entries of Θ^* to be $\frac{1}{\sqrt{s}}$, while other entries are 0, hence Θ^* is (exactly) s-sparse. Sub-Gaussian and heavy-tailed noise ϵ_k are respectively drawn from $\mathcal{N}(0,\sqrt{\frac{3}{5}})$ and $0.3\cdot t(\nu=6)$. All these parameters specify the model, so we can generate the full data $\{(X_k,Y_k):k\in[n]\}$ for a specific (n,d,s).

Then we apply the one-bit quantization scheme to quantize $\{(X_k,Y_k):k\in[n]\}$ to $\{(\dot{X}_{k1},\dot{X}_{k2},\dot{Y}_k):k\in[n]\}$ in 1-bit QC-CS, or $\{(X_k,\dot{Y}_k):k\in[n]\}$ in 1-bit CS. All parameters are properly set, and we stress that the truncation and dithering parameters for X_k, Y_k are different. For instance, in sub-Gaussian 1-bit QC-CS we use dithering noise $\Lambda_k\sim\mathrm{uni}\big([-\gamma_Y,\gamma_Y]\big)$, $\Gamma_{kj}\sim\mathrm{uni}\big([-\gamma_X,\gamma_X]^d\big)$ with $\gamma_X\neq\gamma_Y$. After the data quantization, we can solve the proposed convex programming problems to obtain the estimator $\widehat{\Theta}$. We track the ℓ_2 norm error $\|\widehat{\Theta}-\Theta^*\|_2$ and report the mean value of 15 independent runs.

Algorithm. Note that the convex programming problems (3.17), (3.26) and (3.30) share a common formulation

$$\widehat{\Theta}\in\mathop{\arg\min}_{\Theta\in\mathbb{R}^d}\ \frac{1}{2}\Theta^T\widehat{\mathbf{\Sigma}}_{\mathbf{1}}\Theta-\widehat{\Sigma}_2^T\Theta+\lambda\|\Theta\|_1,\tag{E.1}$$

where $\widehat{\mathbf{\Sigma}}_{\mathbf{1}}$ is positive semi-definite, $\widehat{\Sigma}_2\in\mathbb{R}^d$. Here, we use alternating direction method of multipliers (ADMM) to solve (E.1), and the convergence of our algorithm is guaranteed since the variable is divided into two blocks [45]. For more details of ADMM, we refer readers to the survey paper [17]. We now invoke the framework of ADMM and show the iterative formula. Divide $\Theta\in\mathbb{R}^d$ into $M,Z\in\mathbb{R}^d$, (3.17) is equivalent to

$$\mathop{\arg\min}_{M,Z\in\mathbb{R}^d}\ \frac{1}{2}M^T\widehat{\mathbf{\Sigma}}_{\mathbf{1}}M-\widehat{\Sigma}_2^TM+\lambda\|Z\|_1,\ \text{s.t. } M=Z.$$

By introducing the multiplier $\Upsilon\in\mathbb{R}^d$, the augmented Lagrangian function reads

$$\frac{1}{2}M^T\widehat{\mathbf{\Sigma}}_{\mathbf{1}}M-\widehat{\Sigma}_2^TM+\lambda\|Z\|_1+\Upsilon^T(M-Z)+\frac{\rho}{2}\|M-Z\|_2^2.$$

Minimizing (M,Z) alternatively and updating Υ via gradient ascent give the iteration formulas

$$\begin{cases}M_{t+1}=(\widehat{\mathbf{\Sigma}}_{\mathbf{1}}+\rho\cdot\boldsymbol{I_d})^{-1}(\widehat{\Sigma}_2+\rho\cdot Z_t-\Upsilon_t)\\ Z_{t+1}=\mathcal{S}_{\lambda/\rho}(M_{t+1}+\rho^{-1}\cdot\Upsilon_t)\\ \Upsilon_{t+1}=\Upsilon_t+\rho\cdot(M_{t+1}-Z_{t+1})\end{cases}\tag{E.2}$$

that updates (M_t,Z_t,Υ_t) to $(M_{t+1},Z_{t+1},\Upsilon_{t+1})$. In (E.2), we define $\mathcal{S}_\beta(x)=\mathrm{sign}(x)\max\{0,|x|-\beta\}$ if $x\in\mathbb{R}$, and then let $\mathcal{S}_\beta(\cdot)$ element-wisely operate on vectors. This is known as the soft thresholding operator.

E.3 Low-rank Matrix Completion

Simulation Details. We simulate low-rank matrix completion with exactly low-rank matrix $\mathbf{\Theta}^*$. The $d\times d$ rank r underlying matrix $\mathbf{\Theta}^*$ is generated by the formulation $\mathbf{\Theta}^*=\frac{\mathbf{\Theta_l}\mathbf{\Theta_r}}{\|\mathbf{\Theta_l}\mathbf{\Theta_r}\|_{\mathrm{F}}}$, where entries of $\mathbf{\Theta_l}\in\mathbb{R}^{d\times r}$ and $\mathbf{\Theta_r}\in\mathbb{R}^{r\times d}$ are i.i.d. drawn from $\mathcal{N}(0,1)$. Furthermore, $\mathbf{\Theta}^*$ with different (d,r) are controlled to possess comparable spikiness $\alpha(\mathbf{\Theta}^*)=\frac{d\|\mathbf{\Theta}^*\|_{\max}}{\|\mathbf{\Theta}^*\|_{\mathrm{F}}}$. While the covariate is specified to be $\boldsymbol{X_k}=e_{k(i)}e_{k(j)}^T$ with $(k(i),k(j))\sim\mathrm{uni}([d]\times[d])$ (4.2), we test both sub-Gaussian noise and heavy-tailed noise. Specifically, sub-Gaussian or heavy-tailed ϵ_k are i.i.d. copies of $\mathcal{N}(0, \frac{1}{400})$ or $\frac{1}{250} \cdot \big(\frac{1}{\sqrt{3}} t(\nu = 3)\big)$, respectively. Here, $\frac{1}{\sqrt{3}} t(\nu = 3)$ is the Student's t distribution with 3 degrees of freedom and variance rescaled to 1. Following these parameters, the full data $\{(\boldsymbol{X_k}, Y_k) : k \in [n]\}$ are obtained from the model $Y_k = \big\langle \boldsymbol{X_k}, \boldsymbol{\Theta^*} \big\rangle + \epsilon_k$. The responses are processed by the one-bit quantization scheme and quantized to one-bit $\dot{Y}_k$, then solving the convex programming problem (4.6) gives the estimator $\widehat{\boldsymbol{\Theta}}$. In (4.6), we set $\alpha^* = \|\boldsymbol{\Theta^*}\|_{\max}$ and properly tune λ so that it balances the data fidelity and low-rank structure. We track the Frobenius norm error $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta^*}\|_{\mathrm{F}}$ and report the mean value of 15 independent trials.

Algorithm. We similarly apply ADMM to solve (4.6), and first separate variable $\boldsymbol{\Theta}$ to be two blocks $\boldsymbol{M}, \boldsymbol{Z} \in \mathbb{R}^{d\times d}$. Define $\mathbb{1}'(E)$ to be the indicator function widely used in optimization, i.e., $\mathbb{1}'(E) = 0$ if E happens, $\mathbb{1}'(E) = \infty$ otherwise. Then, we can move the max-norm constraint to objective and obtain the equivalent program

$$\mathop{\arg\min}_{\boldsymbol{M},\boldsymbol{Z}\in\mathbb{R}^{d\times d}} \ \frac{1}{2n}\sum_{k=1}^{n} \big(\big\langle \boldsymbol{X_k}, \boldsymbol{M}\big\rangle - \gamma\cdot \dot{Y}_k\big)^2 + \mathbb{1}'(\|\boldsymbol{M}\|_{\max} \le \alpha^*) + \lambda\|\boldsymbol{Z}\|_{\mathrm{nu}}, \ \text{s.t. } \boldsymbol{M} = \boldsymbol{Z}.$$

Let $\boldsymbol{\Upsilon} \in \mathbb{R}^{d\times d}$ be the multiplier, we have the augmented Lagrangian function

$$\frac{1}{2n}\sum_{k=1}^{n} \big(\big\langle \boldsymbol{X_k}, \boldsymbol{M}\big\rangle - \gamma\cdot \dot{Y}_k\big)^2 + \mathbb{1}'(\|\boldsymbol{M}\|_{\max} \le \alpha^*) + \lambda\|\boldsymbol{Z}\|_{\mathrm{nu}} + \big\langle \boldsymbol{\Upsilon}, \boldsymbol{M} - \boldsymbol{Z}\big\rangle + \frac{\rho}{2}\|\boldsymbol{M} - \boldsymbol{Z}\|_{\mathrm{F}}^2.$$

Some additional notations are necessary before presenting the algorithms. Let $\mathcal{I}_{ij} = \{k \in [n] : \boldsymbol{X_k} = e_i e_j^T\}$, then we define $\boldsymbol{J_1} = [\boldsymbol{J_1}(i,j)], \boldsymbol{J_2} = [\boldsymbol{J_2}(i,j)] \in \mathbb{R}^{d\times d}$ as

$$\boldsymbol{J_1}(i,j) = \sum_{k\in\mathcal{I}_{ij}} \gamma\cdot\dot{Y}_k\ , \quad \boldsymbol{J_2}(i,j) = \sum_{k\in\mathcal{I}_{ij}} 1 = |\mathcal{I}_{ij}|.$$

We define $\mathcal{P}_\Omega(\cdot)$ to be the projection onto $\Omega \subset \mathbb{R}^{d\times d}$ under Frobenius norm. Let $\mathbf{1}$ be the all-ones matrix with self-evident size, and ./ represents the element-wise division between two matrices of the same size. Furthermore, we introduce the soft thresholding operator $\mathcal{S}_\beta(\cdot)$ for a matrix $\boldsymbol{A}$ that admits singular value decomposition $\boldsymbol{A} = \boldsymbol{U\Sigma V^*}$, where the singular values of $\boldsymbol{A}$ are arranged in the diagonal matrix $\boldsymbol{\Sigma}$. Based on $\mathcal{S}_\beta(x) = \mathrm{sign}(x)\max\{0, |x| - \beta\}$ for $x \in \mathbb{R}$, we define $\mathcal{S}_\beta(\boldsymbol{A}) = \boldsymbol{U}\mathcal{S}_\beta(\boldsymbol{\Sigma})\boldsymbol{V^*}$ and let $\mathcal{S}_\beta(\cdot)$ element-wisely operates on the diagonal matrix $\boldsymbol{\Sigma}$. Now, one can derives the ADMM iteration formulas as

$$\begin{cases} \boldsymbol{M_{t+1}} = \mathcal{P}_{\|\boldsymbol{M}\|_{\max}\le\alpha^*}[(n\rho\cdot\boldsymbol{Z_t} + \boldsymbol{J_1} - n\boldsymbol{\Upsilon_t})./(n\rho\cdot\mathbf{1} + \boldsymbol{J_2})] \\ \boldsymbol{Z_{t+1}} = \mathcal{S}_{\lambda/\rho}(\rho^{-1}\cdot\boldsymbol{\Upsilon_t} + \boldsymbol{M_{t+1}}) \\ \boldsymbol{\Upsilon_{t+1}} = \boldsymbol{\Upsilon_t} + \rho\cdot(\boldsymbol{M_{t+1}} - \boldsymbol{Z_{t+1}}) \end{cases}. \tag{E.3}$$